The Recent Excitement in the Coding Problem

F. H. C. Crick
Medical Research Council Laboratory of Molecular Biology
Cambridge, England

I. Introduction

A. The Nature of the Problem

The Sequence Hypothesis states that the amino acid sequence of a protein is determined by the sequence of nucleotides in some particular piece of nucleic acid. The evidence in favor of this is now very considerable, and will not be reviewed here. It is not unreasonable to hope that this relationship may be a simple one and that the sequence of the four bases in the nucleic acid can be thought of as a simple code for the amino acid sequence.

This problem—the exact sequence of bases that determines each of the twenty amino acids found in proteins—is known as the "coding problem." Recently there have been dramatic developments in this field and it now seems possible that the code will be found within a comparatively short time. This review deals critically with the recent progress and discusses the general nature of the genetic code as we can glimpse it today.

B. The Mechanism of Protein Synthesis

The actual mechanism of protein synthesis will not be considered here in any detail. We assume that the main site of synthesis is the ribosomes; that the genetic message is conveyed there by a special species of RNA known as messenger RNA (see Lipmann's article in this volume), which is usually made by copying one of the strands of the double helix of DNA (see Hurwitz and August's article in this volume); and that the amino acids are conveyed to the ribosomes by a special family of RNA molecules known as "soluble RNA" or "transfer RNA" (here called sRNA), which act as "adaptors" and carry each amino acid to its proper site. It is known that the sRNA molecules are specific and that there is at least one type for each amino acid. Each amino acid is joined onto its own sRNA by a special protein—the activating enzyme—in a way that is fairly well understood (Hoagland, 1960).

C. The "Standard" Bases and Amino Acids

The four standard bases in DNA, adenine, guanine, cytosine, and thymine, will be denoted by their initial letters (A, G, C, and T, respectively), and those in RNA by A, G, C, and U, where U stands for uracil. Hypoxanthine will be denoted by H (not by I) and xanthine by X. Thus, polyinosinic acid—the polyribonucleotide all of whose bases are hypoxanthine—will be called poly H.

In fact, there are more than four bases found in DNA and RNA, and more than twenty amino acids in proteins. There are several reasons for ignoring the rarer bases:

(1) Some nucleosides, such as pseudouridine, are only found in types of RNA that do not act as messenger RNA.
(2) No correlation has yet been discovered between any unusual base and any particular amino acid.
(3) All the unusual bases found either in genetic nucleic acid or in messenger RNA are capable of forming one of the two standard base pairs. No suggestion of a third base pair has been found. Thus it is difficult to see how the unusual bases could carry additional "information" (in the technical sense of the word) although they may, of course, *express* information already carried by the base sequence in a different form.

The reasons for ignoring the rarer amino acids are:

(1) All those found in genuine proteins appear to be simple derivates of one of the standard amino acids, e.g., phosphoserine. It is thus a reasonable guess that the modification may be made after the synthesis of the polypeptide chain.
(2) Their distribution is very uneven. For example, hydroxyproline (in animals) is found only in collagen and related molecules.

Nevertheless, there is a possibility that in some cases a rare amino acid may be incorporated using the coding mechanism. For example, it would be interesting to know if there is a special sRNA for hydroxyproline.

There is now general agreement on the most likely set of standard amino acids. This set of twenty (given in Table I) includes asparagine, glutamine, and cysteine, but not cystine. It is perhaps worth noting that the list has remained unchanged since it was first drawn up about 9 years ago.

It has recently been claimed by Zubay (1962) that aspartic and glutamic acid should be removed from the list, as he believes that they are actually incorporated

TABLE I
THE STANDARD SET OF TWENTY AMINO ACIDS[a]

Alanine	Leucine
Arginine	Lysine
Asparagine	Methionine
Aspartic acid	Phenylalanine
Cysteine	Proline
Glutamic acid	Serine
Glutamine	Threonine
Glycine	Tryptophan
Histidine	Tyrosine
Isoleucine	Valine

[a] In alphabetical order.

as asparagine and glutamine (respectively) and then deamidated. This is unlikely for a number of reasons, and further evidence will be needed before his suggestion can be accepted.

In addition, there must be a code of some sort for a "space," or, more likely, for "begin chain" and "end chain."

Curiously enough, a number of proteins have an acetyl group attached to the amino end of the polypeptide chain. The list includes the proteins of several RNA viruses, horse heart cytochrome c, one chain of bovine fibrinogen, α-MSH, and ovalbumin. One wonders whether possibly this may be a way of separating polypeptide chains in those cases where perhaps more than one protein is coded on the same messenger RNA. It would be worth looking for a special sRNA that might perform this function, though it is of course always possible that the acetyl group is added by a special enzyme after the rest of the chain has been synthesized.

Whether a separate language is needed for other purposes, such as control mechanisms, is not known.

D. Glossary of Terms

These definitions are not meant to be precise. They are merely a guide to the current usage. The meaning is usually fairly clear for simple codes, but may become ambiguous for more complicated ones.

1. The Coding Ratio

This is the ratio of bases to amino acids for a suitably long message. For a fully overlapping code the coding ratio is 1. It is often assumed that the coding ratio is an integer, though this need not be the case.

2. The Codon

This is a group of bases that code for one amino acid. In simple codes a codon is a fixed number of consecutive bases, e.g., in a "triplet" code

it is three consecutive bases, but it is possible to conceive codes in which, for example, some codons consist of two bases and others of three. Again it is not certain that the bases making up a codon are adjacent on the polynucleotide chain.

3. Overlapping

In an overlapping code a given base forms part of several codons. For example, in a fully overlapping triplet code one codon may consist of the bases at positions n, $(n+1)$, and $(n+2)$, and the next codon of those at $(n+1)$, $(n+2)$, and $(n+3)$. In a *partly* overlapping triplet code adjacent codons would be at positions n, $(n+1)$, and $(n+2)$ and $(n+2)$, $(n+3)$, and $(n+4)$. In a nonoverlapping code adjacent codons do not overlap.

4. Colinearity

The amino acid sequence and the equivalent sequence of bases on the nucleic acid are said to be "colinear" when the order of amino acids along the polypeptide chain is the same as the order of the corresponding codons along the polynucleotide chain.

5. A Universal Code

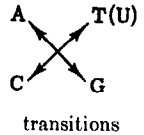
A code is universal if, throughout nature, any particular codon has the same meaning.

6. Degeneracy

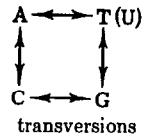
A code is said to be nondegenerate if there is only one codon for each amino acid. Otherwise it is said to be degenerate. In a highly degenerate code almost all the possible codons correspond to amino acids. A code is said to be "logically" degenerate if one can deduce the pattern of degeneracy from a simple rule.

7. Transitions and Transversions

A transition is a change of one purine into the other purine, or one pyrimidine into the other pyrimidine. In a transversion a purine is put in the place of a pyrimidine, or vice versa. Thus we have:



transitions



transversions

8. Sense, etc.

A codon corresponding to an amino acid in the "wild type" gene is referred to as "sense." If such a codon is charged (by mutation) to a

codon for another amino acid, then it is called "mis-sense." If it is changed to a codon that does not stand for an amino acid, it is called "nonsense." If such a change completely destroys the function of the gene, the codon is called "absolute nonsense."

It is not yet known whether all nonsense is absolute nonsense.

9. Ambiguous and Unique

A codon that stands at the same time for more than one amino acid is said to be ambiguous. A code, all the codons of which are unambiguous, is said to be unique. If some of the codons are ambiguous, the code is said to be partly ambiguous.

II. General Questions

A. Is the Code Overlapping?

The first fairly precise code, suggested by Gamow, was of the overlapping type. Gamow (1954) argued that this was plausible for stereochemical reasons. Adjacent amino acids in a polypeptide chain can be no more than 3.7 Å apart, so that, in order to be joined together, they must be brought fairly close to each other. *Adjacent* nucleotides are about the same sort of distance apart—the exact distance depends upon which part of the nucleotide is measured. The bases tend to stack with their planes 3.4 Å apart, but two adjacent phosphates (or two sugars) are usually 5–6 Å apart. However, for a nonoverlapping triplet code the relevant distance is not between adjacent nucleotides but between every third nucleotide. In the usual sort of structure this distance is likely to be 10 Å, or greater. A simple picture of the assembly process would place adjacent amino acids too far apart to be joined together.

We realize now that there are several other ways of getting round this difficulty but at the time it seemed that a fully overlapping code would be rather a neat solution.

Any fully overlapping code will impose restrictions on the amino acid sequence, since the linear density of information is not the same in the two languages, there being at each step only four choices for a base, but potentially twenty choices for an amino acid. Thus the latter choice must necessarily be restricted. It was easy to disprove the code actually suggested by Gamow. This was of the overlapping triplet type and it was fully and logically degenerate. Apart from the fact that it failed to specify the direction of the polypeptide chain, it could be shown that it would not code for the known sequence of insulin.

A search was next made to see if there were any restrictions on amino

acid sequence. This work will not be reviewed here in detail. It suffices to say that no obvious restrictions have been found and that probably any amino acid can neighbor any other, though the data are still too scanty to exclude restrictions on the rarer amino acids, such as tryptophan.

At about this time the adaptor hypothesis was first being considered. This hypothesis (see the account in Hoagland, 1960) states that the amino acid does not find the correct place on the template by specifically adsorbing to it, but is carried there by a special molecule, named an adaptor, that itself can fit onto the right places on the template and to which the amino acid is joined by a special enzyme. In modern biochemical terms the amino acid is joined by a special activating enzyme to a molecule of sRNA (the adaptor), which carries it to the template RNA.

This hypothesis has recently been confirmed by the brilliant experiment of Chapeville *et al.* (1962) who have shown that once cysteine has been attached to its own sRNA it is incorporated by poly U,G as if it were cysteine even though it is chemically turned into either alanine or cysteic acid after being attached. In other words, the amino acid goes where the sRNA directs and has no further control over its own destination.

The importance of this hypothesis, as was realized at the time, is that there is no obvious reason why a particular amino acid should be coded by any particular codon. How the codons and amino acids are associated depends on the exact structure of each activating enzyme and each molecule of sRNA, and could, for all we know, arise from historical accident. This permits codes to be degenerate in a great variety of ways. However, by an ingenious argument, Brenner (1957) was able to show that the nearest-neighbor data ruled out *all* possible fully overlapping degenerate triplet codes. At that time he assumed the code was universal and therefore lumped together the data from all species; now that so many more sequences have been determined, the argument could probably be applied to the nearest-neighbor data from mammals alone, or even from one particular mammal.

However, recent work has made it unlikely that *any* simple overlapping code can be correct. This comes from a study of the changes in amino acid sequence produced by mutation. In a fully overlapping code the change of a base will alter several adjacent amino acids. In particular cases, because of degeneracy, there may be fewer changes, but usually an alteration to a single base in a fully overlapping triplet code will change three adjacent amino acids.

Now all the data suggest that the typical change from a mutation

affects a single amino acid. For example, this is so far true of all the abnormal hemoglobins studied in detail. However, the most useful results have been obtained by treating the RNA of Tobacco Mosaic Virus (TMV) with nitrous acid, and then determining the alteration in sequence (if any) produced in the TMV protein (see Section IV,A for references). Again the typical change affects a single amino acid. No case has yet been reported of a change of two *adjacent* amino acids.

This evidence makes overlapping codes rather unlikely, though it is possible, with ingenuity, to imagine a partly overlapping code that will fit the facts. Such a code has in fact been suggested by Wall (1962) and is discussed in Section VII,C,5.

If the code is nonoverlapping there must be some way of deciding how the bases are grouped into codons, since the RNA structure itself has no obvious "commas" that could do this. Of course a particular base might function as a comma, though this does not seem very likely. At one time it was thought that perhaps the code might be of the "comma-less" type, and could be read in one way only. It can be shown that with four bases this can be done without imposing restrictions on the amino acid sequence if the number of amino acids does not exceed twenty (Crick *et al.*, 1957).

At the moment a comma-less code seems very improbable for a number of reasons. For example, with such a code one cannot easily explain the large variation in base composition of the DNA of different microorganisms. Polyuridylic acid should not act as a messenger, and other results of the cell-free system cannot be explained (see Section VII,C,1).

B. Is the Message Read from One End?

The best evidence that this is so comes from recent genetic studies on the A and B cistrons of the *r*II locus of bacteriophage T4 (Crick *et al.*, 1961). There is no direct evidence that either of these two genes codes for a polypeptide chain, but there is indirect evidence [the ambivalent mutants of Benzer and Champe (1961)] that suggests they do.

The argument turns on the nature of the mutations produced by acridines such as proflavine. Acridines act in this way only on the phage-infected cell, not on free phage, and the action does not need light. It is suspected that the mistake occurs during either replication or recombination, from the acridine slipping in between adjacent base-pairs and forcing them apart (Lerman, 1961). It is postulated that the mutants produced come not from the *alteration* of a base (as in the case of nitrous acid) but from the deletion or addition of one or more of the bases. Mutants of this type, which includes many of the spontaneous mutants

of the *r*II locus, have certain characteristics in common. They are usually nonleaky (that is, the activity of the gene appears to be completely lacking, rather than merely reduced); they cannot be reverted by base-analog mutagens, such as 5-bromouracil or 2-aminopurine; when they do revert, either spontaneously or in the presence of acridines, they usually do so by making a second mutation not too far from the first one. By this method, many such mutants in the B1 and B2 segments of the B cistron have been isolated. It is found that they fall into two classes, called for convenience + and —. For the purposes of explanation, the + class can be considered to have an extra base, and the — class to have one base too few.

By genetic recombination, these mutants can be put together in one gene in pairs. Pairs of the type (+ +) or (— —) show no function, but pairs of the type (+ —) behave rather like the wild type gene, provided the two mutants are not too far apart. How far this can be depends upon whether the + is on the left or the right of the — on the genetic map. We thus have the interesting situation that two mutants, either of which separately will completely destroy the function of the gene, will, when combined within the same gene, allow the gene to function.

All this can be neatly explained by assuming that the message is read from one end in groups of a fixed size, and that in this part of the gene small alterations to the message may not make much difference to the function. Let us suppose that the message is read off, from a fixed starting point, in groups of three bases at a time. The addition of a base will throw the reading out of phase, and beyond that point the message will be completely wrong. Hence such a mutant will have no function. Similarly for the deletion of a base. However, for a double mutant of the type (+ —) the message will only be incorrect between the two mutations, and provided this distance is not too great this may not alter the function of the gene.

The strongest evidence that the message is read from one end comes from the study of a mutant (the deletion 1589) which appears to have joined the A and B cistrons together. This mutant has lost a part of both the A and B cistrons. In spite of this it still shows some B cistron activity. Normally any mutant that maps wholly within the A cistron does not affect the activity of the B cistron, and vice versa. This suggests that there is something between the two cistrons that makes them distinct. It was argued that this small separating region would have been lost in the mutant 1589, and that, since there was now nothing to keep the two cistrons apart, they should be joined together. This proved to be the case, since a mutant of the acridine type in the A cistron, when combined in the same gene as 1589, now destroyed the function of the

B cistron. This has been true for all seven acridine mutants so far tested. Base analog mutants or deletions, on the other hand, when combined with 1589, allow the B function to be expressed in about half the cases tested (for further results see Benzer and Champe, 1962).

It thus seems very likely that the message is read from a fixed point, probably from one end of the cistron. However, in biochemical terms we have two readings: the reading of the messenger RNA from the DNA, and the reading of the protein from the messenger RNA. Although there does not appear to be any very direct evidence it is more reasonable to assume that the latter process is the one disturbed in acridine mutants.

This genetic evidence is very compatible with the biochemical evidence, reported earlier by Bishop *et al.* (1960), Dintzis (1961), and Goldstein and Brown (1961), that the amino acids are assembled into the polypeptide chain in a linear order starting at the amino end.

C. The Coding Ratio and the Size of the Codon

The obvious way to find the coding ratio is to find the size of the piece of nucleic acid coding a protein of known length. Unfortunately it is as yet very difficult to do this with any precision. The usual method is to estimate the *fraction* of the total DNA of the organism that constitutes a particular gene by comparing the length of the gene on the genetic map with the total length of the map. Such an estimate can be only approximate for a variety of reasons. It suffices to say that, in the few cases for which the data is available, the results come to small numbers between, say, 2 and 10.

The recent experiment of Bautz and Hall (1962), in which they appear to have purified the messenger RNA for a particular gene, suggests another possible approach, but here too the method is not likely to lead to a result one can trust without a lot of careful work.

A quite different method comes from the genetic work on the acridine-type mutants of the rII locus of phage T4 described in the last section. Whereas two of these mutants, of like sign, when combined in the same gene, still give the mutant phenotype, *three* such mutants, if close enough together, combine to give a phenotype resembling the wild type (Crick *et al.*, 1961). This obviously suggests that the coding ratio is three. While this is the most likely interpretation, the evidence would also be compatible with it being any multiple of three, though subsidiary observations make this less likely. Unfortunately, more elaborate theories are also not completely eliminated. For example, a code in which the common amino acids had codons of three bases, while a few of the rarer ones had codons of only two bases, might still be compatible with the results

reported so far, though it would be difficult to reconcile them with a predominantly doublet code of the type suggested by Roberts (1962).

If the code is nonoverlapping (and there are no dummy bases between codons) the average size of the codon will equal the coding ratio. If the code is overlapping, or partly overlapping, it will be greater than the coding ratio. At the moment we have no evidence on the codon size apart from that described above for the coding ratio and the absence of overlapping, which together suggests that each codon consists of three bases.

There is one argument frequently used about the probable size of the codon that seems to me to be quite without foundation. Physicochemical studies have shown that the binding of a trinucleotide to a polynucelotide is rather weak. Therefore, it is argued, a codon must have more than three bases.

There might have been some point in this argument at the time when comma-less codes were being considered, since such codes could allow the sRNA to attach to the template well ahead of the growing point. All the more recent work suggests, however, that the synthesis proceeds one step at a time, and that the sRNA need only attach at the actual growing point. This could have a very special environment, and indeed one would certainly expect to find there a protein catalyzing the actual chemical act of polymerization. Thus there is no reason why this protein should not assist, unspecifically, in strengthening the binding of the sRNA, and, as far as I can see, even a codon with one base would suffice from this point of view. After all, DNA appears to be synthesized, in a highly specific manner, one base at a time.

D. Is the Code Universal?

The four bases and the twenty amino acids seem to be essentially the same throughout nature but it does not follow that the code relating them is necessarily everywhere the same. Of course, if each of the amino acids were in some way structurally related to the triplets that represent it [as Woese (1962) appears to believe] then we should certainly expect the code to be universal, but so far no one has been able to suggest any plausible way in which they could be stereochemically related. Indeed it was this very difficulty which led to the adaptor hypothesis (see Section II,A).

It used to be argued that once a complete code had arisen it would be very difficult to change since any alteration in the meaning of a triplet would produce changes in almost every protein in the organism and would thus in all probability be lethal. This argument would lose

some of its force if the code were degenerate and if one particular triplet were used very infrequently. For instance, in an organism having a DNA very high in G and C the triplet AAA would tend to be rare. Under these circumstances mutations in the sRNA and the appropriate activating enzyme, which would allow it to change its meaning, might not be lethal. If, in the course of further evolution, the base ratio of the DNA of such an organism drifted back from the extreme value to one having more A and T, then the triplet AAA might become fairly common.

All this supposes that at some very primitive stage all organisms had the same code. This might have happened if the system arose only once, though it is not obvious that the code should have evolved by a whole series of distinct additions to a simpler system without diversification occurring. In any case, it is very unsatisfactory to have to argue about events that took place a very long time ago under quite unknown conditions, or about very unlikely events that might conceivably have occurred anywhere in biological time. All such arguments, in the last analysis, amount to saying that, since the twenty amino acids are universal and have not changed, one expects the same of the code. This argument is plausible, but not convincing by itself.

Fortunately, in the last year or so, experimental evidence has accumulated suggesting that the code may well be universal. The method used is to synthesize a particular protein in a cell-free system, using some of the components from one species and some from another. Thus, von Ehrenstein and Lipmann (1961) synthesized hemoglobin using ribosomes from rabbit reticulocyte combined with sRNA (with radioactive leucine attached) from *E. coli*. The autoradiographs of the "fingerprints" of the tryptic digests suggest that the hemoglobin made is very similar to, if not identical with, normal rabbit hemoglobin. This experiment has since been successfully repeated (von Ehrenstein, Weisblum, and Benzer, personal communication) with sRNA from both yeast and *Micrococcus lysodeikticus* (which has a DNA with a high GC content).

Assuming that the sRNA has the role usually attributed to it, this certainly suggests that the code for mammals is fairly similar to that for microorganisms.

More recently it has been shown by Tsugita *et al.* (1962) that a protein very similar to TMV protein can be made by adding TMV RNA to a cell-free system from *E. coli*. A similar experiment on the protein of the RNA phage f2 has been reported by Nathans *et al.* (1962).

It has been shown by Signer *et al.* (1961) that the alkaline phosphatase made by an *E. coli* episome in a cell of *Serratia marcescens* is very similar to, if not identical with, that made in cells of *E. coli*, and quite distinct from the alkaline phosphatase made by *S. marcescens* itself. The techniques used were starch electrophoresis and fingerprints of trypsin digests. *E. coli* DNA has 50% (G + C) whereas *S. marcescens* DNA has 58%. They conclude "the type of enzyme synthesized by the cell appears to be determined solely by the genetic material and not by the cellular environment." A similar experiment, using the enzyme β-galactosidase, has been carried out by Jacob in collaboration with Stainer and Condamine, and briefly mentioned by Jacob (1961).

All these experiments suggest that the code may well be universal, but they might also be compatible with codes that, while mainly universal, differ slightly from one organism to another.

Fortunately, it should soon be possible to test the matter in a more direct manner by adding synthetic polynucleotides to cell-free systems from different organisms. A start has already been made by Arnstein *et al.* (1962), who have shown that the stimulation of phenylalanine incorporation by poly U also occurs in a cell-free system from rabbit reticulocytes. [See also Weinstein and Schechter (1962) and Maxwell (1962).]

III. The Cell-Free System

A. Experimental Results and Interpretations

The biochemical details of this system will be mentioned here only very briefly. The components are usually obtained from *E. coli*.

The system generally consists of washed ribosomes, sRNA, and a "supernatant" fraction. It is usually supplied with amino acids, GTP and ATP (and usually an ATP-generating system). Mercaptoethanol, or some similar compound, is usually present and the concentration of Mg^{++} is critical; a typical value is 15m*M*. RNA or various synthetic polynucleotides may be added. Very little net synthesis takes place and, to study amino acid incorporation, one or more of them is made radioactive. At the end of the incubation (15 minutes at 37° is often used), the "protein" is precipitated, for example with hot trichloroacetic acid, washed, and counted.

This system, developed by the studies of many workers, is described in detail by Matthaei and Nirenberg (1961), who refined it. They confirmed earlier reports (Tissières *et al.*, 1960; Tissières and Hopkins, 1961) that amino acid incorporation was partly inhibited by DNase and showed that, while the *initial* rate was unaltered, the subsequent incorporation was greatly reduced. This suggested that the DNA had been acting as a template for the synthesis of mRNA, which in turn was acting as a template for protein synthesis. They, therefore, added RNA from various sources, and found that the RNA from tobacco mosaic virus greatly stimulated amino acid incorporation [a similar stimulation by the RNA of turnip yellow virus was found by Ofengand and Haselkorn

(1962)]. In addition, they tried adding various polynucleotides and made the spectacular discovery that polyuridylic acid (poly U) produced an enormous increase in the incorporation of phenylalanine, and apparently of no other amino acid. Moreover, the product had the rather unusual solubility properties expected of polyphenylalanine (Nirenberg and Matthaei, 1961).

The incorporation needed the usual components—the ribosomes and the supernatant fraction—and was inhibited in whole or in part by puromycin, chloramphenicol, and RNase, but hardly at all by DNase. The presence or absence of the other amino acids made no difference to the incorporation of phenylalanine. In a subsequent paper (Nirenberg *et al.*, 1962), they showed that phenylalanyl sRNA was an intermediate in the synthesis.

This discovery completely revolutionized the biochemical approach to the coding problems, since it offered a relatively simple way of attack by determining which polynucelotides stimulate the incorporation of which amino acids. The initial results, however, were rather negative. No significant stimulation was obtained with poly A, poly H, or poly C, although the latter appeared to produce a small incorporation of proline. Nor was appreciable activity given by a "random" poly A,U.[1] The important observation was made that a mixture of poly A and poly U (which combine to form a triple helix of one poly A and two poly U) did not stimulate the incorporation of phenylalanine. This suggests that single-strandedness is necessary for activity.

On the important question as to whether the synthetic polyribonucleotides act catalytically or stoichiometrically, the present evidence certainly suggests that they work more than once. Matthaei *et al.* (1962) report that under the best conditions a little less than one molecule of phenylalanine is incorporated for each molecule of uracil added as poly U. (This has been confirmed by other workers.) In addition, it has been shown (though not necessarily on the same incubation mixture) that much of the added poly U is rapidly broken down shortly after being added to the system, whereas phenylalanine incorporation goes on for some time (Nirenberg; Boye, both personal communications). This, taken with the fact that the code is unlikely to be overlapping, makes it very improbable that the poly U is acting stoichiometrically.

The most extensive work on the incorporation stimulated by polymers of mixed composition has been done by Ochoa and his colleagues (Lengyel *et al.*, 1961, 1962; Speyer *et al.*, 1962a,b). Using polynucleotide phosphorylase, they synthesized polyribonucleotides containing only

[1] The comma in poly A,U indicates that the proportions of A and U are unspecified.

two or three of the four bases. In all cases one of the bases was uracil, which was present in excess in the incubation mixture. Such polymers naturally incorporate large amounts of phenylalanine, and it was hoped that this would make the polypeptides produced rather insoluble and thus easily recoverable by acid precipitation. Each polynucleotide was tested against all of the twenty amino acids, by adding one amino acid in a radioactive form and all the nineteen others unlabeled. The additional incorporation (above the background observed when no polynucleotide was added) was then calculated on a molar basis, and expressed as the ratio of the molar incorporation of phenylalanine. Their results show that such polymers give substantial incorporation for phenylalanine and six other amino acids (about 20–25% of phenylalanine), some incorporation (between 3–10% of phenylalanine) for ten others, small incorporations for aspartic acid (2½%) and glutamic acid (1½%), and no observable counts for glutamine. A "false" incorporation reported for threonine was about 6% of the phenylalanine.

The observed incorporations were then interpreted as follows. It was assumed:

(1) that the code is based on triplets;
(2) that the composition of the polynucleotide is the same as that of the incubation mixture used to produce it;*
(3) that, to a first approximation, the sequence is random.

It was thus possible to calculate the expected frequency of any particular triplet. It was tacitly assumed that usually only one triplet stood for each amino acid, and the effect of possible nonsense triplets was not seriously considered. Only the triplet UUU was assigned to phenylalanine, and triplets were assigned to the other amino acids on the basis of the ratio of their incorporation compared with that of phenylalanine. Naturally only the composition and not the order of the triplets can be obtained by this method. By this means Ochoa and his colleagues were able to allot a triplet to all the twenty amino acids except glutamine. These triplets are listed in Table III in Section III,E. Two triplets were suggested for asparagine, three for leucine, and a second doubtful one (UCC) for threonine—otherwise one triplet was given for each amino acid.

These results are of very considerable interest, though in a subsequent paper (Speyer *et al.*, 1962b), the authors indicate that "the possibility that . . . the code may be more extensive than is apparent at present . . . should be kept in mind."

Similar results for fourteen of the twenty amino acids have been

* See Addendum.

TABLE II

AMINO ACID INCORPORATION INTO PROTEIN STIMULATED BY RANDOMLY MIXED POLYNUCLEOTIDES[a,b]

Polynucleotide	UA	UC	UG	UAC	UGC	UGA	Composition of coding units[c]
Base-ratio	U = 0.87 A = 0.13	U = 0.39 C = 0.61	U = 0.76 G = 0.24	U = 0.834 A = 0.050 C = 0.116	U = 0.341 G = 0.152 C = 0.502	U = 0.675 G = 0.291 A = 0.034	
Probability of triplet relative to phenylalanine (UUU = 100%)	UUU—100 UUA— 13 UAA— 2.2 AAA— 0.3	UUU—100 UUC—157 UCC—244 CCC—382	UUU—100 UUG— 32 UGG— 10.6 GGG— 3.4	UUU—100 UUA— 6.0 UAA— 0.4 AAA— 0.02 UUC— 13.9 UCC— 1.9 CCC— 0.3 UAC— 0.8 AAC— 0.05 ACC— 0.12	UUU—100 UUG— 46.2 UGG— 21.0 GGG— 1.0 UUC—147 UCC—218 CCC—332 UGC— 68.1 GGC— 31.7 GCC—101	UUU—100 UUG— 43 UGG— 19 GGG— 8.1 UUA— 5.1 UAA— 0.26 AAA— 0.01 UGA— 2.2 GGA— 0.1 GAA— 0.01	
Amino acid							
Phenylalanine	100	100	100	100	100	100	UUU...
Arginine	0	0	1.1	0	**49.3**	2.9	UCG...
Alanine	1.9	0	0	1.0	**40.4**	0.9	UCG...
Serine	0.4	**160**	3.2	3.6	**170**	2.3	UUC.. + UCG...
Proline	0	**285**	0	0	**188**	0	UCC...
Tyrosine	**13**	0	0	8	1.0	**8.6**	UUA...
Isoleucine	**12**	1.0	1.0	4.8	5.4	**8.4**	UUA...
Valine	0.6	0	**37**	0.4	**29.8**	**75**	UUG...
Leucine	4.9	**79**	**36**	5.1	**157**	**44**	UUC.. + UUG...
Cysteine	4.9	0	**35**	0	**5.4**	**46**	UUG... or UGG
Tryptophan	1.1	0	**14**	0	1.6	**23**	UGG...
Glycine	4.7	0	**12**	0.5	**9.7**	**15**	UGG...
Methionine	0.6	0	0	0.6	1.5	**8**	UGA...
Glutamic acid	1.5	0	0	1.2	0.44	**6.2**	UGA...
Lysine							UAA...(?)

[a] Taken from Matthaei *et al.* (1962).

[b] The figures in the main part of the table represent the incorporation of any amino acid compared to phenylalanine incorporation expressed as percentages (mμmoles amino acid incorporated/mμmoles phenylalanine incorporated × 100). Boldface figures refer to the polynucleotide containing the nucleotides necessary to stimulate the incorporation of a given amino acid. The reproducibility of the above percentage figures was ±3.

[c] Sequence of nucleotides in a coding unit is not specified.

reported by Matthaei *et al.* (1962). The essential features are reproduced in Table II, taken from their paper. This should enable the reader to grasp the principle used in allocating triplets and to form some idea of the agreement between theory and practice. It should be noted that this agreement might be improved by allocating further triplets, especially to phenylalanine. For example, the results for poly U,G,C would fit better if one gave (UUC)[2] to phenylalanine in addition to UUU, and also gave both (UCC) and CCC to proline. This illustrates some of the difficulties of interpretation that arise.

The triplets allotted by these authors agree with those of Ochoa's group, except that they do not confirm (UUA) as one of the triplets for leucine, nor the doubtful (UCC) for threonine. Matthaei *et al.* (1962) are not clear whether cysteine is represented by (UUG) or (UGG). They suspect that, in addition to (UUC), serine may also be represented by (UCG).

It has recently been reported that a poly C,A containing either 85% or 63% C will stimulate the incorporation of proline and some threonine and histidine (Bretscher and Grunberg-Manago, 1962). The poly C,A was carefully analyzed and contained less than 1% U (and about 0.8% G). It has also been found that a poly A,G,C stimulates the incorporation of several amino acids (Bretscher, personal communication). However, the possibility has not been eliminated, in either of these cases, that some deamination may not take place in the incubation mixture. With this reservation, it appears that a polymer need not contain U to work in the cell-free system, though polymers with an excess of U certainly seem to work best.*

B. Criticisms of the Experiments

There are so many criticisms to be brought against this type of experiment that one hardly knows where to begin. The major criticism of the work is that there is no measurement of the composition of the polynucleotides used. The assumption that the polynucleotides have the same composition as the incubation mixture from which they are made is known to be invalid in some cases, especially for mixtures involving G. Reluctantly, one must put to one side most of their results for polynucleotides containing G until their composition has been measured.*

A second major criticism is that small amounts of incorporation

[2] UUC denotes a triplet with the bases in that order. (UUC) denotes a triplet in which the order of bases is unknown.

* See Addendum.

should be looked on with suspicion. Two groups of workers (Matthaei *et al.*, 1962; Bretscher and Grunberg-Manago, 1962) have reported that, in addition to phenylalanine, poly U stimulates the incorporation of some leucine, usually between 5 and 10% according to the latter workers, who have gone to some lengths to show that this is not a simple artifact due to impurities, etc. Moreover, it has been shown (Nirenberg, personal communication) that, if precautions are taken to remove all traces of phenylalanine, the incorporation of leucine in the presence of poly U may be considerable (up to 50% of the usual phenylalanine incorporation).

Whatever the reasons for this effect, it implies that small amounts of incorporation should be mistrusted. That this is not needless precaution is shown by the case of threonine. Polymers containing various amounts of U and C will incorporate only small, though usually variable, amounts of threonine. Since the poly U,C first tested was mostly uracil, the triplet (UCC) was expected to be rare, and thus this small incorporation of threonine was ascribed to (UCC). However, a poly U,C containing mostly C, in which the triplet (UCC) should be common, incorporates little, if any, threonine (Bretscher and Grunberg-Manago, 1962; Nirenberg, personal communication).

These results underline the necessity of measuring *several* polynucleotides with related composition, and of carefully measuring their composition and their purity, before allocating triplets. This has so far been done in only a few cases (for poly U,C and poly U,G) (Nirenberg, personal communication) and, at the time of writing, the results have not been published.

C. Criticism of the Interpretation

The remaining criticism concerns the methods used to allocate triplets to amino acids. It should be remembered that the biochemical experiments themselves give no justification for a code consisting entirely of triplets, and indeed some of the results (for example, those for poly U,C) would fit equally well a code consisting mainly of doublets. The main evidence for a wholly triplet code is from the genetic work on the *r*II gene of phage T4, and this can only be accepted with reservations. However, the assumption that all codons are triplets is at the moment by far the most likely, and it is reasonable to proceed on this basis.

The same cannot be said for the assumption that the code is completely or almost completely nondegenerate; this is a favorite of authors not closely acquainted with the problem. This is discussed in detail later. The method used by Ochoa and his colleagues to allocate triplets would work well if the code were nondegenerate, or at any rate only slightly degenerate, and if the effect of nonsense triplets could be

ignored, either because the mechanism skipped over them in some way or, alternatively, if they led to spaces in the transcription, that is, to short polypeptides, which nevertheless were sufficiently insoluble to be precipitated by acid. For example, since only two of the three possible triplets of (UUC) have been allocated to amino acids, it is tacitly implied that one of them is nonsense. Thus, a polymer with the composition of 5 U's to 1 C would lead to a nonsense triplet about one time in six, and the average length of peptide (assuming no skipping) would be about 5 residues, assuming that the base sequence is random.

The other theoretical difficulty in allocating codons concerns polymers that do not stimulate incorporation appreciably, such as poly C. At first sight this suggests that CCC is a nonsense codon, but another alternative is that poly C may not work because it has secondary structure, for which there is some evidence in that it shows some hyperchromic change with temperature, unlike poly U, which shows none. In fact, there is a suspicious correlation between polynucleotides with secondary structure and inactivity in the cell-free system. In particular, it has been shown (Nirenberg and Singer, personal communication) that if polynucleotides containing U and various amounts of G are tested in the cell-free system, there comes a point, as the proportion of G is increased, when the incorporation activity starts to decline. This is just the point where physical chemical studies show the onset of secondary structure.

To give an example, the triplet (UCG) has been allotted to arginine, but there is no strong reason why this amino acid should not also be given the triplet (GGC). Even if a poly G,C does not act in the cell-free system, it would have to be shown that this failure did not come from secondary structure, from a nonsense triplet (such as perhaps GGG), or from a failure to recover the peptides owing to too high a solubility, before it could be shown that (GGC) does not code arginine.

The methods used by Ochoa's group, then, if carefully applied, can lead to the plausible allocation of *certain* codons. For example, the quantitative studies of Nirenberg and his colleagues (unpublished) make it highly probable that (assuming triplets) one of the (UUC) codons represents serine and one of the (UUG)'s represents valine. But to attempt to identify other codons for serine and valine leads to many difficulties. "Errors" of incorporation, lack of randomness of sequence, failure to recover certain peptides, and uncertainties about secondary structure all make the allocation of further triplets very precarious.

D. On Random Sequences

In interpreting the incorporation stimulated by synthetic polynucleotides in the cell-free system, it is usually assumed that the sequence of these polymers is random, or approximately so. There is very little experimental data on this and theoretical studies (Simha and Zimmerman, 1962) suggest that what there is may not be too reliable. Unfortunately, it involves a considerable amount of rather careful work to establish the point even in the most favorable case, which is that of a polyribonucleotide containing only two bases, one a purine and the other a pyrimidine.

However, there are some limitations to the frequency of base pairs, triplets, etc., even when the sequence is very far from random. Thus *any* sequence whatever (of infinite length) of two letters X and Y presents the doublet XY as often as the doublet YX. Moreover, it is easy to prove the more general theorem that the occurrences of XY_n equal those of Y_nX. For example, XYYY occurs as often as YYYX (the $n = 3$ case). Similarly for the pair X_nY and YX_n.

To give a concrete example, let us suppose that the triplets UCC and UCU were attributed to leucine, while CUC and CUU were given to serine. Then, if the incorporation of leucine is observed to be less than that of serine, this cannot be attributed in a simple way to a nonrandom sequence of the poly U,C since it is easy to prove that for any (infinite) sequence

$$\text{UCC} + \text{UCU} = \text{UC} = \text{CU} = \text{CUC} + \text{CUU}$$

The same result is obtained if leucine is given UUC and CCU and serine has CUU and UCC.

Thus some discretion must be used in attributing *all* variations in abundance to nonrandomness of sequence.

E. The Ochoa Triplets Assessed

As the set of triplets proposed by Ochoa and his colleagues (Speyer *et al.*, 1962b) has been widely quoted, I have undertaken the thankless task of trying to assess the probability of their correctness, using all the evidence available at the time of writing. I have considered *only* the triplets suggested by Ochoa, and not all those that more recent work has made equally plausible [for example, (ACC) for threonine]. I have assumed that the code is based entirely on triplets, and have not taken any serious note of the possibility of ambiguous codons. The triplets set out in Table III are classified under the headings "probable," "possible," and "doubtful," but there are obviously borderline cases.

Thus, the decision to consider (UGG)-Try as probable and (UGG)-Gly as possible is a marginal one. It has also been very difficult to decide what to accept for leucine. Nevertheless, the table may have some use, if only to produce a more critical attitude to the allocation of codons.

F. Polynucleotides Containing Unusual Bases

A start has been made (Basilio *et al.*, 1962) on the study of polymers containing bases other than the standard four. Thus a poly U,H incorporates in a way very similar to a poly U,G, although the activity of the former appeared to be rather less than that of the latter. That H acts like G has been confirmed by other groups (Nirenberg; Bretscher and Grunberg-Manago, both personal communications). No poly U,X has been synthesized, but the action of nitrous acid on poly U,G, which would be expected to turn G into X, drastically decreased the activity of the polymer.

TABLE III
THE OCHOA TRIPLETS ASSESSED

Probable			
Phe	UUU	100%[a]	Certainly coded by U's alone
			Codon may be ambiguous
Ser	(UUC)	25%	N,[b] B[c]
Pro	(UCC)	8%	N, B, certainly high in C
Ileu	(UUA)	20%	(N, B)
Tyr	(UUA)	25%	(N, B)
Val	(UUG)	20%	N
Leu	(UUG)	12%	N
Try	(UGG)	5%	N
Possible			
Ala	(UCG)	3%	N
Arg	(UCG)	3%	N
AspN	(UAA)	7%	
Cys	(UUG)	25%	N, observed only half expected incorporation (non-random sequence?)
Gly	(UGG)	4%	(N)
Lys	(UAA)	3%	(N)
Met	(UGA)	4%	(N)
Thr	(UAC)	9%	Difficult to assess
Leu	(UUA)	14%	(N, B) but amount less than expected. Difficult to assess
Leu	(UUC)	20%	(N, B) but amount less than expected. Difficult to assess
Doubtful			
Asp	(UAG)	2½%	No evidence it has to contain U
Glu	(UAG)	1½%	No evidence it has to contain U
His	(UAC)	3%	No evidence it has to contain U
GluN	(UCG)[d]	0%	No evidence at all from cell-free system
AspN	(UAC)	—	No direct evidence for this
Thr	(UCC)	6%	? by Ochoa. Results of N and B make it unlikely

[a] The percentages are taken from Table 2 of Speyer *et al.* (1962b). They show the incorporation, on a molar basis, compared with that of phenylalanine.

[b] N = Nirenberg and colleagues } without parentheses = quantitative support
[c] B = Bretscher and colleagues } with parentheses = observation confirmed

[d] Guessed from mutagenic change in TMV protein.

Both these results might have been expected from similar results obtained with the DNA polymerase and the RNA polymerase, and are broadly compatible with what is expected from considerations of base-pairing, though it has never been completely clear why xanthine always acts quite so badly. Preliminary work has also started on the synthesis of polymers that contain bases unable to form the standard base pairs (Lengyel, personal communication).

Haschemeyer and Rich (1962) have tested poly T (polyribothymidylic acid) and found that it does not stimulate the incorporation of phenylalanine. They do not state whether they tested other amino acids. They surmise that this may be because poly T may form a multistranded helix with itself (Rich, unpublished observations).

G. Summary

At this point it may be useful to see what general conclusions can be drawn from the work on the cell-free system. The positive conclusions are:

(1) The amino acid incorporation produced by any particular polymer is quite characteristic, and, in general, different for polynucleotides of different composition, though it may not be completely specific, since some variation in incorporation is observed.
(2) The four letters A, U, C, and G appear distinct in their meaning, since the incorporations produced by poly U, poly U,C, poly U,A, and poly U,G are all quite different from each other. Thus, two-letter codes, such as one in which U = G and A = C, are eliminated. As might have been expected, however, H appears to act very like G.
(3) The incorporation is unlikely to be a complete artifact, since biochemically it appears to require the same steps as genuine protein synthesis (involvement of sRNA and ribosomes, the effects of inhibitors, etc.), and there is some correlation with the mutagenesis results (Section IV) and with Sucoka's (1961a,b) results on over-all composition (Section V,4). This is too good to be simply chance.
(4) Assuming a triplet code, about one third of the codons suggested

by Ochoa and his co-workers are likely to be right, and most of the rest may be correct, though evidence for several of them is inadequate (e.g., aspartic, glutamic, and histidine).

(5) The synthetic polynucleotides probably work more than once, that is, catalytically rather than stoichiometrically.

The following negative results should be kept in mind:

(1) It has yet to be shown by a direct method that the amino acids incorporated are in peptide linkage.*

(2) The size of the polypeptide chains produced and their end groups have not yet been determined.

(3) The direction of the code is not yet known; that is, it is not known whether the end of the polynucleotide having a 3′ hydroxyl corresponds to the amino or to the carboxyl end of the polypeptide chain.*

(4) There is no direct biochemical evidence on the size of the codon.

(5) The total number of codons standing for amino acids is not yet known even approximately. It seems likely that there are at least two codons for leucine. There is no convincing evidence to suggest that all codons must contain U and fairly good evidence that some do not.*

(6) It is not known which codons are nonsense, nor, even approximately, how many there are of them, nor what the system does when it encounters such a codon.

IV. Amino Acid Changes from a Single Mutation

A. Experimental Data

Many cases are known of species or strain differences in the amino acid sequence of a protein and it may well be that many of these differences have been produced by the change of a single base of the nucleic acid. This, however, is a dangerous assumption to make at this stage, especially if the two proteins differ in a number of places along the polypeptide chain.

The changes considered here are those in which the two sequences differ only in a single amino acid or can be considered to be very closely related biologically. Most of the data comes from the studies on either the protein of TMV or on human hemoglobin.

The protein of TMV has been extensively studied by Wittmann

*See Addendum.

TABLE IV
AMINO ACID REPLACEMENTS OBSERVED IN TMV PROTEIN

Change observed				
from	to	Peptide[a]	Cases	Author
		From HNO_2		
Arg	Gly	III(61)	1	T[b]
Arg	Gly	IX(134)	1	T
Arg	Gly	X(122)	1	T
Asp+[c]	Ser	I	2 + 1	W[d] + T
AspN	Ser	VI(73)	1	W
AspN	Ser	XI	1	W
Asp+	Gly	I	2	W
Asp	Gly	IV	2	W
Asp	Ala	I	4	W
Asp+	Ala	X	2	T
GluN	Val	I	2	W
Glu	Gly	VIII(97)	1 + 1	W + T
Ileu	Val	I(24)	1	W
Ileu	Val	X	2	W
Ileu	Met	I	1	W
Leu	Phe	I	1	W
Pro	Ser	IV(63)	3	W
Pro	Leu	I	1 + 1	W + T
Pro	Leu	XII(156)	1	T
Ser	Phe	I	1	W
Ser	Phe	XI(138)	3 + 2	W + T
Ser	Phe	XII	1	T
Ser	Leu	I	1	W
Ser	Leu	III(55)	1	W
Thr	Ileu	I	2	W
Thr	Ileu	III(59)	2	W
Thr	Ileu	X	4	W
Thr	Met	VIII	3	W
Thr	Ser	I	2	T

TABLE IV (*Continued*)

Change observed				
from	to	Peptide[a]	Cases	Author
From spontaneous mutation				
Asp^+	Ala	I	2	W
AspN	Arg	I	1	W
AspN	Lys	XI	1	W
Ileu	Thr	X	1	W
Leu	Phe	I	1	W
Ser	Phe	I	1	W
From *N*-bromosuccinimide				
Arg	Gly	II(+6)	1	T
Asp^+	Ser	I	3	T
Asp^+	Ser	?	1	T
Ileu	Thr	?	1	T
Pro	Leu	I	4	T
From dimethylsulfate				
Arg	Gly	?	1	T
Pro	Leu	I	3	T
Ser	Phe	XI(138)	1	T

[a] The roman numerals indicate that the change is in a particular tryptic peptide. The number in parentheses shows (where it is known) the position on the polypeptide chain, numbering from the amino end. The multiple changes observed by Tsugita have been omitted as it is suspected that they may have arisen from contamination.

[b] T = Tsugita (1962).

[c] Asp^+ stands for Asp or AspN.

[d] W = Wittmann (1962).

(1962) and Tsugita (1962). Most of the changes have been produced by nitrous acid, many of them in what appears to be a single step from the common wild-type strain of TMV. Others have occurred "spontaneously," or have been produced by *N*-bromosuccinimide (NBSI) or by dimethylsulfate (DMS). They are listed in Table IV. The "spontaneous" changes observed in "abnormal" human hemoglobins are listed in Table V.

Two cautionary remarks should be made about such data. First, although a particular mutation may appear to have been made by a particular mutagen, one can never be completely certain of this, since there is always a background of spontaneous mutation; in the work on

TABLE V

AMINO ACID REPLACEMENTS IN HUMAN HEMOGLOBIN[a]

Position[b]	From	To	Name of hemoglobin
α-chain			
16	Lys	Asp	Hb I
30	Glu	GluN	Hb $G_{Honolulu}$
57	Gly	Asp	$Hb_{Norfolk}$
58	His	Tyr	Hb M_{Boston}
68	AspN	Lys	Hb $G_{Philadelphia}$
β-chain			
6	Glu	Val	Hb S
6	Glu	Lys	Hb C
7	Glu	Gly	Hb $G_{San\ Jose}$
26	Glu	Lys	Hb E
63	His	Tyr	Hb $M_{Saskatoon}$
63	His	Arg	Hb MZ_{urich}
67	Val	Glu	Hb $M_{Milwaukee}$
121	Glu	GluN	Hb $D\beta_{Punjab}$–$D\gamma$

[a] For references see Smith (1962a) or Perutz (1962).

[b] The numbering is from the amino end.

TMV, precautions have been taken to make this background as low as possible. Again, even if chemical studies have revealed the usual action of a mutagen, rarer side reactions cannot be ruled out. Thus it is unwise to put too much weight on changes (for example, Ileu → Met) that have only been observed once.

Second, it is important to realize that one is seeing mutations through a phenotypic screen. Only those changes producing a tolerably acceptable protein will be picked up, and many possible changes may be expected to be lethal or sublethal. Again, the actual technique used may select only certain changes. All those observed so far in the human hemoglobin produce an alteration in charge and, indeed, most of them were originally detected by electrophoretic differences between the mutant and the wild-type protein. Nor is it surprising, for example, that for TMV protein no changes have been observed starting from either methionine or histidine, since those amino acids do not occur in the wild-type protein.

Amino acid changes in the A protein of the tryptophan synthetase of *E. coli* have been reported by Yanofsky and his collaborators (Yanofsky *et al.*, 1961; Helinski and Yanofsky, 1962). The changes so far published concern a glycine in the wild type that in one mutant (A-23) has become

arginine and in another (A-46) has been replaced by glutamic acid. Both mutants were produced by ultraviolet light. Genetic studies have shown that A-23 and A-46 can give wild recombinants, though at a low rate. This suggests that the two mutants are altered in different bases in the same codon.*

B. Mutagenic Patterns

The main changes expected by the action of nitrous acid are:

$$C \rightarrow U$$
$$A \rightarrow H = G$$

since nitrous acid turns adenine into hypoxanthine, which is then expected to pair in the same way as guanine. No effect would be expected on uracil, and the xanthine produced by the action of nitrous acid on guanine will either pair in the same way as guanine, or, more likely, be lethal.

Assuming, then, that the two changes set out above are the only ones that can occur, it is possible to work out the pattern of changes produced by nitrous acid on the sixty-four triplets. A typical pattern is shown in Fig. 1. This figure contains all those triplets containing only U's and C's.

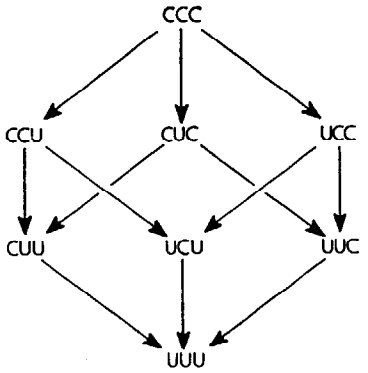


FIG. 1. The changes expected to be produced by nitrous acid for eight of the sixty-four triplets. This figure is referred to in the text as a cube. The complete mutagenic pattern for all the sixty-four triplets consists of eight cubes (see Wittmann, 1962, who calls them octets).

It can be seen that the arrows can be placed to fall on the edges of a cube, with a triplet at each corner, and this pattern will therefore be

* See Addendum.

called a cube. The sixty-four triplets fall onto eight distinct cubes, each having at the top a triplet containing only A's and C's and at the bottom a triplet having only U's and G's.

The corresponding pattern for a doublet code is the diamond, shown in Fig. 2. The sixteen doublets fall onto four distinct diamonds. Notice that if we consider the class of all the *transitions* (that is changes of one

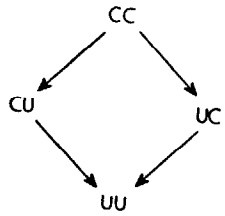
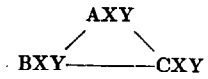


FIG. 2. The changes expected to be produced by nitrous acid for four of the sixteen doublets. This figure is referred to in the text as a diamond. The complete mutagenic pattern for all sixteen doublets consists of four diamonds.

purine for another purine, or one pyrimidine for another pyrimidine) we get exactly the same patterns except that the arrows now run in both directions, since the transitions are:

$$C \leftrightarrow U$$
$$A \leftrightarrow G$$

We now consider the class of nondegenerate codes. Then it is easy to see that it is impossible to produce a triangle by transitions alone. The only way to obtain a triangle is to have the situation:



which is only possible if we have one transition and two transversions. Since nitrous acid is believed to produce only transitions, one should not obtain a triangle from a nondegenerate code with nitrous acid.

The importance of mutagenic patterns becomes apparent when one studies codes that are logically degenerate, since the pattern expected is independent of the allocation of triplets to amino acids. Thus, provided only one set of triplets is given to each amino acid, the mutagenic pattern for transitions for the code proposed by Woese (1962) can easily be shown to consist of two cubes and two diamonds. No triangles can occur.

C. Interpretation

The combined results for all the nitrous acid mutants of TMV are shown in Fig. 3. It is striking that in no case do we have an arrow

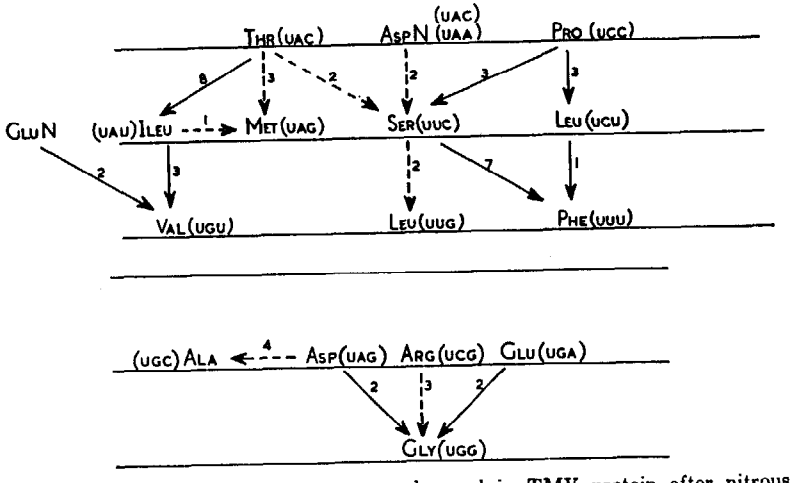


FIG. 3. The amino acid replacements observed in TMV protein after nitrous acid treatment. The numbers show the number of occurrences. The triplets are those proposed by Lengyel *et al.* (1962). Solid arrows show agreement with the expected changes, dotted arrows disagreement.

running in *both* directions between two amino acids. If we have the change A → B, then the change B → A is not observed. This is exactly what we expect if the code is nondegenerate and if nitrous acid acts as expected.

It will be noticed that leucine occupies two places in the figure. Otherwise, one obtains a triangle for the changes between proline, serine, and leucine. We also have a triangle for threonine, isoleucine, and methionine, but the change Ileu → Met has only been observed once.

We must now compare the results with Ochoa's triplets obtained from the cell-free system. To facilitate this the amino acids have been placed on three lines, the top line having amino acids whose triplets contain AA, AC, or CC, the second line those with one A or one C, and the bottom line for those with neither. Leucine has been put in two places, one corresponding to (UUC) and one to (UUG). It is seen that all the arrows run downwards, except for two that are horizontal. No arrow runs upwards.

Again, this is just what we expect. Incidentally, this strongly suggests that the actual RNA strand of TMV (and not its complement) acts as messenger RNA, in line with the experiments of Tsugita *et al.* (1962) on the synthesis of TMV protein in the cell-free system.

Next we see which arrows can be accounted for by *any* change of a single base. This is possible for all the changes except AspN → Ser, and even this is possible if we allow AspN to be (UAC). However, when we allow only the changes expected of nitrous acid (A → G, C → U), we find that only eight arrows fit (leaving aside GluN → Val, which was deduced from this evidence and not from data on the cell-free system), whereas seven arrows do not. In other words, the detailed agreement is rather poor. We must conclude either that nitrous acid often does not act as expected or that the Ochoa triplets are inadequate to explain the mutagenesis. Of course the situation can be improved by invoking other triplets, such as (UAC) for AspN and (UCC) and (UCG) for Ser, and so forth, but it is difficult to know where to stop.

The changes observed in human hemoglobin are set out in a similar way in Fig. 4. Again, if we allow AspN to be (UCA) all the arrows

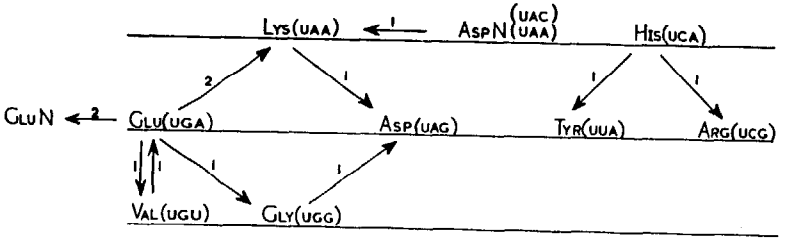


FIG. 4. The amino acid replacements observed in human hemoglobin. The numbers show the number of occurrences. The triplets are those proposed by Lengyel *et al.* (1962).

correspond to the change of only one base in the Ochoa triplets. However, it should be remembered that the evidence from the cell-free system for the triplets for Glu, Asp, and His is very weak, and for GluN is nonexistent.

My own view is that the general agreement between the mutational changes and the results of the cell-free system shows convincingly that the cell-free system is not a complete artifact, and is related to genuine protein synthesis. The detailed disagreement suggests that the present allocation of triplets is only partly correct.

V. Further Experimental Evidence

A. Over-all Composition of Protein and DNA

One of the major dilemmas in the coding problem has been the wide range of DNA base composition observed, especially for microorganisms (Lee *et al.*, 1956; Belozersky and Spirin, 1958). The content of G + C goes from 72% for *Micrococcus lysodeikticus* to 35% for *Bacillus cereus*, and as low as 25% for strains of *Tetrahymena pyriformis*. Moreover, the distribution of base-composition of the DNA "molecules" of any one organism is generally a rather narrow unimodal one (Sueoka *et al.*, 1959; Rolfe and Meselson, 1959). It was expected, on general grounds, that the proteins of these organisms would not differ enormously in amino acid composition, and this has indeed been shown to be the case by Sueoka (1961a,b).

Sueoka has studied the over-all amino acid composition of the proteins of about twenty microorganisms. Cell wall material was discarded but ribosomes were included. However, Sueoka showed that for most amino acids the results were not very sensitive to the conditions of growth, suggesting that bias from some proteins being common while others are rare was not too great. If the ribosomes are removed, the composition of the rest of the soluble proteins shows the same correlations with DNA composition, although the absolute content of the basic amino acids is less (Sueoka, personal communication).

Sueoka reported results for all fifteen amino acids (tryptophan was not measured) and for (Glu + GluN) and (Asp + AspN). By and large the amounts found did not vary too much from one organism to another, but for certain amino acids he found a strong trend with DNA base composition. Thus there was usually about twice as much alanine in an organism with a DNA having 72% G + C compared with one having only 25% G + C.

His results can be classified according to their correlation with G + C content as follows:

− −	Ileu, Tyr, Lys, Phe
−	His
0	Thr, Leu, Val
+	Gly, Pro
++	Ala, Arg

This is not quite the same classification as given by Sueoka (1961a). I have used as a criterion the slope divided by the mean value. On this scale (Glu + GluN) and (Asp + AspN) would be −. I have not included methionine (which would nominally be − − or −) because, unlike the other amino acids, the results are very variable and show little real correlation with base composition.

Too much importance should not be given to this data and especially to the exact shape of the curves. The protein sample may be biased and the DNA base sequence is certainly not completely random. Nevertheless, as Sueoka points out, the correlations do support the idea that the code is rather similar for all organisms, although one wonders about the results for methionine. They also make a nondegenerate triplet code appear unlikely, but are compatible with various degenerate triplet codes, or with a predominantly doublet code (Sueoka, 1961a).

TABLE VI
COMPARISON OF SUEOKA'S RESULTS WITH THE OCHOA TRIPLETS[a]

− −	Phe	(U)UU
	Tyr	(U)UA
	Ileu	(U)UA
	Lys	(U)AA
−	His	UAC
	Ser	UUC
0	Thr	(U)AC
	Leu	(U)UC
		(U)UG
	Val	(U)UG
+	Gly	UGC
	Pro	UCC
++	Ala	(U)CG
	Arg	(U)CG

[a] The signs in the first column show the correlation found by Sueoka (1961a) with G + C content of the DNA. The triplets are those suggested by Lengyel *et al.* (1962). The triplets UUA for leucine and UCC for threonine have been omitted.

It is interesting to compare the slopes found with the set of triplets suggested by Ochoa and co-workers. This is done in Table VI. It is seen that excellent agreement is reached if one is allowed to discard one U or not at will.

B. The Fractionation of sRNA

The fractionation of sRNA is important for several reasons. First, it may help to decide how degenerate the code is. Second, it is likely to prove a most useful tool for identifying codons and for comparing them

in different proteins, and in particular in deciding just how close the code is to being universal.

It is not proposed to review here all the work on the fractionation and characterization of sRNA, and consideration is restricted to two cases. Holley and his collaborators (Doctor *et al.*, 1961) fractionated yeast sRNA by countercurrent distribution and find two peaks for leucine activity and a double peak for threonine. The peaks for alanine, valine, tryptophan, and tyrosine appear single, though of course each may consist of a mixture of species. This enabled Weisblum *et al.* (1962) to perform a very significant experiment. They tested, in the cell-free system, the two leucine fractions separately against poly U,G and poly U,C, both of which stimulate the incorporation of leucine, and find that one of the sRNA fractions, when loaded with radioactive leucine, incorporates only with one of these polymers, and the other fraction only with the other polymer. In brief, there appears to be one leucine sRNA for (UUG) and another for (UUC). This is really the first strong evidence for degeneracy, although Berg *et al.* (1962) have also shown chemically that the leucyl sRNA of *E. coli* is of at least two different types. The obvious next experiment is to use the two sRNA fractions, one labeled and the other unlabeled, for synthesis of a definite protein such as hemoglobin, and to show that one sRNA puts radioactive leucine only into certain places in the amino acid sequence and not into others.

Sueoka and Yamane have also fractionated sRNA, this time from *E. coli*, using the more rapid and convenient method of the Lerman-Hershey column (Sueoka and Yamane, 1962). They observed what appear to be multiple peaks for isoleucine, histidine, glutamic acid, leucine, tryptophan, serine, and valine, and possibly for arginine and threonine. The profiles for alanine, glycine, aspartic acid, lysine, methionine, phenylalanine, and tyrosine are not unlike single peaks, though, once again, these peaks may be more heterogeneous than they appear. So far it has not been proved that these multiple peaks, as in the leucine case, are genuine evidence of degeneracy, but it is not unreasonable to expect that some of them are. It is far more difficult to estimate from the available data the *total* number of different types of sRNA, and this must be left to further attempts at fractionation.

By studying the attachment of the amino acid to the sRNA, Berg *et al.* (1961) showed that there are in *E. coli* two distinguishable acceptors for methionine, since an enzyme prepared from yeast attaches methionine to only one of them. Other workers have studied the interaction between the sRNA from one species and activating enzyme from another, and find that in some cases the foreign enzyme works equally well and in others less well or not at all (Rendi and Ochoa, 1961; Benzer and Weisblum, 1961). However, if there are two distinct sites on the sRNA, one for interaction with the activating enzyme and one for base-pairing with the template, we expect, for evolutionary reasons, results similar to those reported, even if the code were universal. This evidence by itself does not prove that the code is not universal. Nor without further experiments can it be assumed that, for example, the two species of methionine sRNA (mentioned above) each code for different codons, though this may well turn out to be the case.

C. The RNA of the Ribosomes

It has been assumed in this paper that the template for protein synthesis is a special RNA whose composition and sequence follows that of one chain (or, less likely, both chains) of the DNA of the cell, and that the rRNA, which accounts for most of the RNA of the cell, is inert and does not act as a template for protein synthesis.

Such a view can always be questioned until such a time as the template for the synthesis of specific proteins is clearly identified, and indeed until some reasonable function is found for the rRNA. It should be noted, however, that, whereas viral RNA and even random poly A,U,C,G, when added to the cell-free system, stimulate amino acid incorporation, the addition of pure rRNA has little or no effect. It is not known whether this is because the rRNA has been damaged, because of its secondary structure, or because it contains many nonsense codons.

The base composition of rRNA, which is relatively constant throughout nature, though possibly slightly correlated with the DNA base composition, has an unusual feature that has been remarked on by Roberts (1962). It is somewhat similar to the base composition deduced from the over-all amino acid composition of the proteins of the cell, translated using the "doublet" code of Roberts, which can be loosely described as the Ochoa triplets with one U removed from each. Woese (1962) has made a similar observation in relation to the code proposed by him.

The reason for this coincidence is obscure. Of course, if the rRNA were really a specific store for activated amino acids, and if the code were a triplet code of the XY· type, we might expect this, but, attractive though the idea is, present evidence does little to support it. This problem must be left for the future.

D. Genetic Information

It has yet to be shown that the gene and the protein it specifies are colinear. It would be surprising if they were not, and experimental proof may not be far away. Meanwhile, there are two proteins for which it has been demonstrated that mutants near one another on the genetic

map produce changes close to one another in the amino acid sequence. The first is the alkaline phosphatase of *E. coli* studied by Rothman, Levinthal, and Garen, and briefly reported by Rothman (1961). The other, which is better documented, is the A protein of tryptophan synthetase of *E. coli* studied by Yanofsky and his collaborators (Helinski and Yanofsky, 1962; Henning and Yanofsky, 1962a). So far no case has been reported that shows good evidence against colinearity, that is, in which mutants near together on the genetic map produce changes far apart in the polypeptide chain, or vice versa.

Much evidence has accumulated, especially from the study of the rII locus of phage T4, on the effect of different mutagens in both the forward and reverse directions, on the characterization of nonsense mutation, on the effect of 5-fluorouracil in producing phenotypic suppression of certain mutants (Champe and Benzer, 1962), and on similar effects produced by suppressor mutants in other genes [ambivalent mutants—Benzer and Champe (1961)]. The last phenomenon has also been studied using the tryptophan synthetase and alkaline phosphatase of *E. coli* (Yanofsky *et al.*, 1961).

This work will not be reviewed here in detail as it would take too much space to describe it adequately and the conclusions, though suggestive, will necessarily be rather uncertain until supported by biochemical studies. However, it should be mentioned that there is evidence that some nonsense codons exist, indicating that the code is unlikely to be completely degenerate, and that a codon of a gene can be nonsense in one host and to some extent sense in another (Benzer and Champe, 1962; Garen and Siddiqi, 1962). This suggests that the code is not completely universal, though it is not easy to see in exactly what way.

VI. Is the Code Degenerate?

A. The Experimental Evidence

The evidence that the code is degenerate is of two types, direct and indirect, and, with one exception, none of it is satisfactory.

The direct evidence comes from the cell-free system. It has been claimed that leucine is represented by (UUC), (UUA), and (UUG). In view of the incorporation of leucine by poly U, this evidence by itself is not very convincing. However, the demonstration (mentioned briefly in Section V,B) that one fraction of sRNA appears to code for (UUG) and the other for (UUC) does suggest that in this case the degeneracy is real.

The original claim that threonine is represented by (UCC) seems unlikely to be correct, but it may perhaps be represented by (UAC) and (ACC). Other claims, such as (UAA) and (UAC) for asparagine, are not properly established. In fact, the use of polymers of known composition but (presumably) random sequence makes it difficult to discover if more than one triplet in the polymer is representing a given amino acid.

There is, however, indirect evidence for degeneracy from these experiments, but this involves assumptions about what the system will do when it comes to a nonsense triplet. Consider the case of polymers containing only U and C. There is general agreement that such polymers incorporate only proline, serine, leucine, and phenylalanine and that the relative amounts, for polymers of widely different ratios of U to C, are compatible with the triplets:

(UCC)	Pro
(UUC)	Ser, Leu
UUU	Phe

although there is some doubt about the exact amount of leucine and a small, somewhat variable, amount of threonine is also incorporated. There are, of course, in a random polymer of U and C eight different triplets, so with the above scheme half the triplets would be unaccounted for. One might expect that in such a situation either the polymer would not work as messenger, or the average length of polypeptide chain produced would be only two or three residues long. A more palatable alternative is that all eight triplets are used, as follows:

CCC	Pro
(UCC)	*Pro*, Ser, Leu
(UUC)	*Ser*, *Leu*, Phe
UUU	*Phe*

It will be seen that the relative amounts incorporated in this scheme will be just the same as for the previous one (represented here by the allocations in italics), so one cannot distinguish between them in this way. The usual argument against the double allocation is that it makes CCC stand for proline, whereas poly C incorporates no amino acids, apart from a trace of proline. The counterargument to this (as discussed earlier) is that poly C probably has some (unknown) secondary structure at neutral pH—in any case, solutions of it are very viscous—and that it is this secondary structure that prevents the incorporation. In favor of this view is the fact that polymers with a lot of C and a little U, or a little A, work rather well and incorporate a lot of proline. This would be unlikely, it is argued, if CCC represented nonsense; on the other hand, the small amounts of either U or A could easily break up the secondary structure.

The matter is clearly an important one, and turns on the behavior of the system when confronted with a nonsense triplet and on the secondary structure of the various polymers, both of which at this stage are unknown.

The other indirect evidence is of two main kinds. There is, first, the genetic argument from the work of Crick *et al.* (1961) on the acridine mutants of the *r*II locus. As stated earlier, a combination (+ −) will work satisfactorily provided the two mutant sites are not too far apart. If, however, they are widely separated, such a combination has the mutant phenotype. It was postulated that this happens because in certain places the shift of reading frame produces an "unacceptable" triplet, possibly a nonsense triplet. The position of such triplets depends upon whether the + is on the left or the right of the −, as theory would suggest. More recent work on the same system (Shulman, personal communication) shows that such an unacceptable site can be made "acceptable" by base-analog mutagens, and enables the position of these sites to be mapped.

A very rough calculation suggests that such unacceptable triplets are not too common, or it would not be possible to make (+ −) combinations unless the two mutations were very close together. It is thus surmised that nonsense triplets are rare, and thus that the code is highly degenerate. However, this assumes that all nonsense triplets prevent the production of any part of the polypeptide chain. It may be that this is only true of some of them, which we may call "absolute nonsense," whereas others, while not standing for an amino acid, may allow part of the polypeptide chain to be produced. For mutants in this region of the B cistron, such a truncated protein might well be active. If this were so, then the data would merely show that absolute nonsense is rare and would not tell us how degenerate the code is.

Thus the whole argument for high degeneracy is rather indirect and involves too many assumptions to be accepted with confidence, although it may well be correct.

The second type of indirect evidence for degeneracy is the wide range of DNA base composition found compared with the rather small range of amino acid composition (see Section V,A). This is most easily explained by certain types of degenerate code, but alternative explanations are not completely eliminated. For example, a large part of the DNA might involve control mechanisms and thus be irrelevant for coding.

Paradoxically, if each of the twenty amino acids were represented by a triplet containing uracil, as proposed by Ochoa and his colleagues, this would be rather good evidence for degeneracy. It seems highly unlikely that all messenger RNA contains large amounts of uracil (Chargaff, 1962; Reichmann *et al.*, 1962). The natural way out of this dilemma is to assume that the triplets so far allocated are only a fraction of the codons actually used.

On the other hand, evidence suggesting that the code is completely nondegenerate is rather slender and is contradicted by the data on leucine. Certainly the evidence from the cell-free system, either published or implied in discussion, does not support it. Moreover, the impression given that all polymers without uracil have been tried and have failed to work is misleading, and there are other reasons, such as secondary structure, that might explain why certain polymers do not work. The advancement, at this time, of a completely nondegenerate universal triplet code would require much special pleading.

B. Possible Types of Degeneracy

Although it seems unlikely that the code is nondegenerate, the nature and the amount of the degeneracy is unclear. However, if it is highly degenerate, it does not seem probable that it is so at random. To construct a randomly degenerate triplet code, imagine that all the sixty-four triplets were put in a box and then drawn out at random giving, say, five to leucine, two to methionine, etc. The main reason against a code of this type, apart from its inherent inplausibility, is that the mutations produced by nitrous acid do not seem to vary sufficiently. It seems more plausible that the triplets for any particular amino acid will be rather alike. This can be formally stated as follows. We call two triplets "connected" when one can be changed into the other by altering one base only. Then the expectation is that all the triplets for, say, leucine will be connected, as for example:

A connected code has the property that the amino acid changes produced by the alteration of a single base are usually somewhat fewer than for a random code, since some base alterations will leave the amino acid unchanged. Moreover, it is easy to invent ways in which such a code could have evolved.

A special subclass of the connected codes is of particular interest. This is those codes that are *logically* degenerate; that is, for which the pattern of degeneracy is given by a simple rule, as in the code suggested by Woese (1962) (Section VII,D). (Actually a code can be logically degenerate without being connected in the above sense; for example, the

combination code of Gamow and Ycas, 1955.) Certain logically degenerate codes, such as Woese's code, reduce very considerably the changes expected from a mutagen such as nitrous acid. Whatever the reason, the changes so far observed to arise from mutation, produced either by nitrous acid or otherwise, appear to be of rather few kinds, though how much the present sample is biased it is almost impossible to say. In fact, many of the changes can be roughly fitted to a doublet code that is also compatible with the data from the cell-free system (Roberts, 1962). A pure doublet code only provides sixteen codons, and does not allow a wide range of alternative base-ratios for the same message. But it does suggest that if the code is really a pure triplet code its degeneracy makes it look at times more like a doublet one.

C. Degeneracy and the sRNA

At first sight it might be thought impossible to construct a triplet code which was highly degenerate using only one molecule of sRNA (with only one recognition site) for each amino acid. This can easily be done, however, if some molecules of sRNA recognize only two bases of a triplet and are indifferent to the third, although the reading of the base sequence moves on three bases at each step. Such a limited recognition can be symbolized by XY·, where the dot indicates that any of the four bases can occur at that position. Naturally, it is only possible to code a maximum of sixteen sets in this way. However, if some molecules of sRNA recognize all three bases, then one easily obtains enough codons. Suppose, for example, that we have six of the amino acids coded by

AAG, AAU, AAC, GGA, GGU, and GGC

(one for each), with AAA and GGG as spaces, and suppose that all the others (except for AA and GG) are of the form XY·, giving fourteen further amino acids, making twenty in all. Such a code need have only one sRNA for each amino acid, yet it is highly degenerate (all triplets stand for amino acids except AAA and GGG) and compatible with a fairly large range of DNA base compositions, since the composition of the third base in the triplet can be freely adjusted in many cases.

I do not feel this code is correct, but it illustrates the point that there may be fewer sRNA molecules than one might surmise at first sight. Moreover, it would not be surprising if the actual code had *something* of this character. It bears obvious resemblances to the code suggested by Roberts (1962), except that it is strictly a triplet code and not a mixture of triplets and doublets (i.e., the coding ratio = 3, and not between 2 and 3 as in Roberts' code). It should be noted that a mixture of codons of the type XY· with some of the type X·Y is likely to lead to ambiguous triplets. As discussed in Section VII,D, one interpretation of Woese's code also calls for only one sRNA molecule per amino acid, but in his case this would imply that the sRNA did not always use the usual base pairs to recognize the template RNA, whereas in the above scheme the standard pairs can be used throughout.

D. Summary

We can summarize the evidence on degeneracy as follows:

(1) It seems highly unlikely that the code is completely nondegenerate. This is contradicted both by data from the cell-free system and from the TMV nitrous acid mutants.

(2) The amount of degeneracy may be small, i.e., perhaps thirty or fewer of the sixty-four possible triplets may stand for amino acids. The evidence from the cell-free system, the amino acid replacement data, and the fractionation of sRNA, could all, at the moment, be compatible with this.

(3) The amount of degeneracy may be very much higher.* This is suggested by the wide range of DNA composition and by the genetic studies. It is not contradicted by the more direct evidence, though this suggests that if the code is highly degenerate it is unlikely to be degenerate at random.

I favor the third alternative.

VII. Theoretical Matters

A. The Order of Bases within a Codon

It is obvious that the study of polynucleotides of random sequences can lead only to the *composition* of the codon and not to the *order* of the bases within a codon. However, it was realized at a very early stage that the amino acid changes observed in one-step mutation could provide some information on the *relative* order within different codons, though not on the *absolute* order. We could arbitrarily call the three positions within a triplet the *a*, *b*, and *c* positions, and describe all the triplets in this same way (that the *a* position in one corresponded to the *a* position in all the others, etc.), and leave to further study the problem of the *actual* order of the *a*, *b*, and *c* positions on the polynucleotide chain, for which there are obviously six alternatives.

Most workers on the coding problem have made preliminary attempts along these lines for their own private use but only a few authors have

* See Addendum.

published them. Zubay and Quastler (1962) sought to deduce the code from amino acid replacements alone, using a method that involved obtaining, by an unspecified technique, a rather poor fit to some doubtful data. They assumed that the code was nondegenerate (which is unlikely), and that aspartic acid was equivalent to asparagine and glutamic acid to glutamine. The latter assumption was later supported by the paper by Zubay (1962) mentioned in Section I,C. It suffices to say, for example, that their (nondegenerate) assignment of UUA to Glu and UGG to Ser can only be reconciled with the results from the cell-free system by rather special pleading.

Two other attempts, by Smith and by Jukes, proceeded on more obvious lines. In both cases, the assignments are based on the triplets suggested by Ochoa and his colleagues. At the time that Smith was working on his first paper (Smith, 1962a), codons had been proposed for only fourteen amino acids. On the tacit assumption that the code was nondegenerate, he was able to deduce the codons for four others from the amino acid replacement data. These were subsequently confirmed by Speyer *et al.* (1962b) but the evidence for three of them can only be classified as doubtful (see Table III). In a second paper, Smith (1962b) produced ordered codons for all twenty amino acids, the only ambiguities being for Cys and Try. He pointed out that "the assumption of degeneracy . . . is unnecessary to explain the amino acid substitutions in proteins already presented." He continued, "There are, however, certain exceptions," which he mentioned. However, he did not explain why, if the code were nondegenerate, he should ignore these mutations and not others; or that if additional codons are necessary, this would make his whole method uncertain. The paper includes a discussion of why all codons contain U, a fact neither supported by the experimental evidence nor likely on general grounds.

Jukes (1962a) based his first code partly on a quite unlikely set of assumptions about a pair of amino acid changes in β-lactoglobulin, which were shortly afterwards shown to be incorrect. A second attempt (Jukes, 1962b) proceeded on now familiar lines, being based on the Ochoa triplets, but allowing a small amount of degeneracy. His assignments, as might be expected, are not the same as Smith's but are not completely dissimilar, having most of the triplets in common. This is not surprising, since they were based on almost identical data and assumptions.

It is clear that these attempts were premature. Apart from the fact that many of the triplets on which they are based are not firmly established, the method can only work if the composition of all (or almost all) codons is known, at least for the amino acids being considered. Even

then, it must be used with discretion, since it is unlikely that *all* the amino acid replacements observed will fit (a few may involve a change of more than one base), and it must be shown that only one solution fits the data very well and that all other solutions fit very poorly.

Nevertheless, it is worth noting that a tolerable fit is obtained even with the present data. The basic reason for this is that, as pointed out earlier in Section VI,B, one can construct a *doublet* code that is compatible with much of the data from the cell-free system and from amino acid replacements (Roberts, 1962), provided one ignores the way nitrous acid is expected to act. My own belief is that this amount of agreement is possible not because the code consists mainly of doublets but because the triplets are not degenerate at random, and because the replacement data, especially for hemoglobin, is at the moment rather limited.

B. Ambiguous Codons

The incorporation of some leucine, in addition to phenylalanine, by poly U, mentioned earlier in Section III,B, suggests that ambiguous codons may exist. The evidence so far presented could be explained as an artifact of the cell-free system, or an effect peculiar to a few strains, though preliminary experiments show a similar effect in a cell-free system from rabbit reticulocytes (Arnstein, personal communication).

There does not seem, unfortunately, any good theoretical reason why it should not be a universal property of the code, provided ambiguity were restricted for a few codons. An ambiguous codon would then be like a weak form of nonsense, or double-talk. Occasionally we should expect to find a single gene producing two amino acid sequences (differing by one amino acid), though natural selection might make this a rare occurrence in wild-type proteins.

There is not likely to be much enthusiasm for this idea but it should not be lost sight of. Theoretically it can be shown that if all amino acids are represented by either XY·, X·Y, or ·XY then one can code up to twenty-four different things and still allow each of them to have two unambiguous triplets (Crick and Watts-Tobin, unpublished). On the present evidence, however, such a code seems unlikely.

C. Unlikely Codes

Although the number of firmly established facts is rather small they are enough to make several codes very improbable.

1. The Comma-less Code (Crick *et al.*, 1957)

There are many members of this family of codes but they all have in common the fact that, of the three cyclic permutations of any codon,

only one should be sense and two nonsense. Thus the "probable" results from the cell-free system (Table III) eliminate comma-less codes. Several other less direct arguments can also be made against them.

2. The Combination Code (Gamow and Ycas, 1955)

In this code all the permutations of any triplet stand for the same amino acid. Thus GUU is the same amino acid as UGU and UUG. Again the results from the cell-free system make this unlikely.

3. Transposable Triplet Codes

There are several types to be considered. The "transposition" is the complementary triplet on the other chain of the DNA, read in the opposite direction. For example, the transposition of UCG is CGA. One old favorite is that both the triplet and its transposition stand for the *same* amino acid. Another suggestion is that, if a triplet is sense, then its transposition is nonsense. In other versions, the complementary sequence is read in the *same* direction, i.e., UCG and AGC are considered a pair.

None of these codes seem very likely, if only because they restrict the variation one can obtain, by degeneracy, in the base composition of the DNA. However, they might be disproved if we had more definite data from the cell-free system. In particular, the number of different amino acids coded by (UUA) and (UAA) should not exceed three. Table III suggests that four or five amino acids may be coded in this way, but only two of these triplets are classed as "probable," so that the matter must remain open for the time being. Another possible group to consider is (UGG) plus (CCA).*

4. Two-Letter Codes

In these codes, the bases were put into two groups. Thus, it was suggested (Sinsheimer, 1959) that the two letters were 6-keto (G = U) and 6-amino (A = C). As already stated, the whole character of the results from the cell-free system make all such codes unlikely.

5. Partly Overlapping Codes

It has been pointed out by Wall (1962) that partly overlapping codes are not completely eliminated by the present data. He considers various possibilities and proposes, in particular, a code having a codon *size* of 4, but a coding *ratio* of 3, so that the first and last base of each codon are shared with the two adjacent codons. He further suggests that perhaps

* See Addendum.

only U and G may occur in these positions, although all four bases are allowed in the two middle places of the codon. Wall rightly points out that the results of Crick *et al.* (1961) show that it is the coding *ratio* (and not the *size* of the codon) that is likely to be 3 and are thus in this respect compatible with his suggestion. Moreover, although the code is partly overlapping, nitrous acid would not produce changes in two adjacent amino acids, as neither A nor C is shared by two adjacent codons. In fact the only changes that would alter two adjacent amino acids are G→U or U→G, and if such transversions were not too common they might not have been picked up in, say, the abnormal hemoglobin, especially if the number of possible alterations were reduced by systematic degeneracy. He also argues that double changes might be lethal more often than single changes, as they may cause more damage to the structure of the protein. For all these reasons the present data on amino acid substitutions is certainly not incompatible with such a code.

However, the *distance apart* of the double or triple mutants observed by Crick *et al.* (1961) to have the wild phenotype is not easily compatible with his code, since the deletion or addition of a base would be expected to lead to a nonsense codon in many cases. The fact that a poly C,A appears to act as a messenger is also against this particular form of his code, since all templates should contain either U or G or both. Nor can it be said that there are any strong arguments in favor of his code. Nevertheless, it remains a possibility that should be kept in mind.

D. Woese's Code

The balance of evidence suggests that the code is highly degenerate but not degenerate at random. A code of this type, which is also logically degenerate, has been put forward on a detailed basis by Woese (1962). He assumed the code was of the nonoverlapping triplet type, and that it should be degenerate in such a way as to account for the large range of DNA base-ratios. He considered codes in which the degeneracy is caused by equating certain bases, depending upon whether the first, second, or third letter of the triplet is being considered. As we have seen from the results on the cell-free system, the four bases are distinct, but it does not follow that they are distinct in all three positions. Such a code gives $a \times b \times c$ different sets of triplets, where a, b, and c are integers $\leqslant 4$. The smallest number greater than 21 obtained in this way is $2 \times 3 \times 4 = 24$. (The next smallest is $3 \times 3 \times 3 = 27$.) That is, there should be one position in the codon with two alternatives, another with three, and at the third all four bases should be distinct. The code he chose can be symbolized as:

$$\begin{bmatrix} \text{U} \\ \text{C} \\ \text{A} \\ \text{G} \end{bmatrix} \quad \begin{matrix} \text{A} = \text{C} \\ \text{G} = \text{U} \end{matrix} \quad \begin{matrix} \text{C} = \text{U} \\ \text{A} \\ \text{G} \end{matrix} \Bigg] = 4 \times 2 \times 3$$

1st 2nd position 3rd

and Woese pointed out that there could be a structural basis to this. The degeneracy in the second position is obtained by having as the two alternatives 6-amino (A or C) or 6-keto (G or U), and in the third position the two pyrimidines, U and C, both have a keto group at the same place.

There is, however, no compelling reason why this particular choice should be made. There are forty-two distinct codes of the type $2 \times 3 \times 4$, not including permutations of the *order* of the letters, and many of these could be given some structural justification. However, if one is restricted to codes allowing a *large* variation of DNA base composition and predicting, for example, that poly U,C will incorporate just four amino acids (rather than two or eight), there are only eighteen codes that need be considered. It is not clear at the moment that any of these codes have any clear advantage over the one proposed by Woese, and they are not considered further here. *No* codes of this type give a mutagenic pattern for transitions containing *triangles* unless one amino acid is allotted to more than one of the twenty-four sets of triplets.

Having by his rule allocated the sixty-four triplets to twenty-four distinct sets, Woese proceeded to determine which amino acid went with which set of triplets, using as a guide the triplets suggested by Ochoa and some of the amino acid replacement data. The solution he proposed is set out in Table VII. In judging Woese's code, one must realize that some of the data he used to derive it may have been misleading and that other allocations are not ruled out. He was able to fit most of the triplets suggested by Ochoa, except that he could not include both cysteine and tryptophan, but only one of them. This springs from the fact that only eighteen of his sets contain U. He was also uncertain whether to exchange the sets allocated to serine with those for leucine, and possibly arginine with alanine and glutamine with aspartic acid.

The agreement is not unimpressive, since it predicts that a poly C,A would incorporate proline and threonine, which was not known when he allocated his triplets.

A code of this type can be considered in two distinct ways. The fact that there is a structural basis for the degeneracy suggests that there may be only *one* sRNA for each amino acid but that it does not always use the usual base-pairs when it sits on the template, so that the same sRNA can combine with more than one triplet. Thus the grouping of the triplets into twenty-four sets would be dictated by some unknown but

plausible structural requirement, although the allocation of any set to a particular amino acid might be a historical accident, assuming the correctness of the adaptor hypothesis.

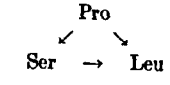
It is not easy to reconcile the experimental data to this point of view since the fact that an amino acid is *not* incorporated by a certain polymer is then difficult to explain away. Moreover one might expect that H (which appears to act like G, as one would predict if the usual base-pairs are used) could act like A in the third position of the triplet, since both have a CH at position 2 in the purine ring.

The alternative point of view, favored by Woese (personal communication), is that there is a *separate* sRNA for each triplet, making, on his allocation, fifty-six types in all. The relationship between the different triplets for one amino acid is then attributed to some unknown stereochemical relationship between each particular amino acid and its codons in the remote past, so that leucine, say, is necessarily associated with certain triplets. That such a general relationship could be discovered seems to be unlikely (in fact the adaptor hypothesis was invented because of this difficulty).

At any event, no such stereochemical relationship has yet been suggested. Woese's code, in this form, does not appear to have an adequate *theoretical* foundation.

Woese's code, as given in Table VII, is not in perfect agreement with the experimental data. For example, the more reliable results for the cell-free system show no incorporation corresponding to (UAU) for serine, or (GCU) for methionine. However, it could be argued that the sRNA for these triplets are missing, or that the base sequence of the polymers is grossly nonrandom. If the three possible (UUG) codons stand for Leu, Cys, and Val, then Phe should not be (UGU), but this argument is a weak one. Again, the apparent incorporation of leucine because of a (UUG) and of histidine because of (ACC) is not accounted for by his code, but it is just possible that there are "errors" of incorporation. In short, the evidence from the cell-free system is not sufficiently precise to make it possible to reject Woese's code with certainty.

Most of the amino acid replacement data will fit his scheme, but no possible allocation of triplets could account for the triangle



found for nitrous acid mutants, unless one of these amino acids is represented by more than one of his sets of triplets, or unless one of these changes arises either from a transversion or from transitions to

TABLE VII

THE CODE PROPOSED BY WOESE (1962)[a]

1.	**UUU**, UUC, UGU, UGC	Phe
2.	UAU, **UCU**, UAC, UCC	Ser
3.	**CUU**, CUC, CGU, CGC	Leu
4.	CAU, **CCU**, CAC, CCC	Pro
5.	**GUU**, GGU, GUC, GGC	Val
6.	**UUG**, UGG	Cys
7.	**GUG**, GGG	Gly
8.	**AUU**, AUC, AGU, AGC	Ileu
9.	**UUA**, UGA	Tyr
10.	**AUA**, AGA	Lys
11.	**UAA, UCA**	AspN
12.	AAU, **ACU**, AAC, ACC	Thr
13.	**GAU**, GCU, GAC, GCC	Met
14.	**AUG**, AGG	Asp
15.	**GUA**, GGA	Glu
16.	**CUG**, CGG	Ala
17.	UAG, **UCG**	Arg
18.	**CUA**, CGA	His
19.	CAG, CCG	—
20.	GAG, GCG	—
21.	CAA, CCA	—
22.	AAA, ACA	—
23.	GAA, GCA	GluN
24.	AAG, ACG	—

[a] The amino acids are assigned according to the experimentally observed coding triplets (which are boldface) of Speyer *et al.* (1962b) and according to amino acid replacement data.

two bases, in his case UAU (Ser) → CGU (Leu), or UCU (Ser) → CUU (Leu). The alteration found in hemoglobin of His → Arg would involve a change of all three bases. However, if arginine were reallocated to set no. 16 (CUG and CGG) only one base would need to be changed, A → G. The mutagenic pattern for nitrous acid mutants (Fig. 3) shows five arrows connected to serine. Again (with the reservations made above) no possible allocation of the sets of triplets can give this result. In fact, it cannot be obtained with any of the $2 \times 3 \times 4$ codes, since it can be shown (Crick, unpublished) that there is no case of a set being connected by transitions to more than four other sets.

The situation will clearly have to be reviewed from time to time as more reliable data becomes available, but the amount of disagreement at the present time is disturbing.

It seems unlikely that this code, or similar codes of the $2 \times 3 \times 4$ type, can be correct. The importance of Woese's code is that it appears to be, in a rough sense, the sort of code one might expect. It is, therefore, of interest to see how far one may go in fitting other types of data to it.

Woese (personal communication) has made the following observations:

(1) Given the approximate amino acid composition associated with an organism, a *range* of nucleic acid compositions can be calculated compatible with it. The actual base composition, in all cases studied, falls within this range. That is, one can account for the large variation in DNA base-ratios in microorganisms, and also the RNA composition of six small viruses, including wild cucumber mosaic virus, containing 40% C.

(2) Woese shows that his code allows one to fit these data on nearest-neighbor frequencies for the bases in DNA from a variety of microorganisms (Josse *et al.*, 1961) rather well if the absolute order of bases is that shown in Table VII. Four of the other five possible permutations of order are a bad fit. He can only do this if the two strands of the DNA of any structural gene do not have identical base compositions (in all cases studied he predicts G > C for mRNA).

(3) Swartz *et al.* (1962) have reported the nearest-neighbor base frequency for the single-stranded DNA virus ϕX174. Woese shows that his code accounts for these frequencies if one assumes that the DNA chain in the virus is the coding strand. He states that one cannot fit the frequencies if the complementary strand is the coding strand.

One may reserve judgement on all this for two reasons. First, in fitting the data, there are a good number of disposable parameters, although Woese shows very convincingly that they are not elastic enough to allow *any* two sets of data to be fitted together. Second, Woese's code has been correlated with Ochoa's triplets, and, as we have seen, these fit Sueoka's data fairly well. Thus, I suspect that many codes of this general sort (that is, with the right kind of degeneracy and fitted to Ochoa's triplets) could be made to fit the data on the base composition and nearest-neighbor data of the nucleic acid. Obviously *small* (illogical) variations in Woese's code would not alter the fit too much, and indeed in one place he argues that this may have happened in the course of evolution. Clearly a final judgement on this point cannot be made until the code is known completely. It would be of interest to see if any of the other $2 \times 3 \times 4$ codes would fit the facts as well.

VIII. General Observations

A. The Nature of the Code

It is clear from what has been said that at almost every point we lack certain knowledge of the genetic code. Nevertheless, the weight of evidence certainly suggests that it is *a nonoverlapping triplet code, heavily degenerate in some semisystematic way, and universal, or nearly so*. Whether this general impression is misleading remains to be seen. However, although not one single codon can be said to be known with certainty, we do know something: one codon for phenylalanine contains U's, one for proline contains C's, and so on. The coding problem has moved out of the realm of rather abstract speculation into the rough and tumble of experimentation.

B. Future Developments

It would be rash to attempt to forecast the exact way in which the subject will develop, but certain trends can be foreseen.

We can distinguish two types of information we need: first, the general nature of the genetic code; second, the distinctive identification of each codon. The general question on which information is most needed is that of degeneracy. Until we have a better idea of the number of codons that stand for amino acids, other evidence will be difficult to interpret.* A likely approach lies in the fractionation of the sRNA. Where in the past the main aim was to obtain the sRNA which accepted one amino acid free from all the others, now we require the separation (or at least the separate labeling) of the different sRNA molecules accepting the *same* amino acid. It does not matter whether our samples are heavily contaminated with sRNA for other amino acids, since we can ignore these by having just the one amino acid radioactive.

These RNA fractions can then be tested in the cell-free system for incorporation with defined polynucleotides, and also into particular positions in defined proteins. Ultimately this technique should allow us to decipher the exact base sequence in a gene in spite of the ambiguity caused by degeneracy.

The other obvious development is the synthesis of polynucleotides with defined or partly defined sequences. The most likely would seem to be either polymers made by polynucleotide phosphorylase and starting with small known primers,* or polynucleotides made chemically, with repeating sequence such as XYXYXY . . . or XYZXYZXYZ. . . . Moreover, there is always the hope that by luck, or from increasing knowledge

*See Addendum.

of the actual mechanism of protein synthesis, it may be possible to use very small polynucleotides, even trinucleotides, in some way to uncover the code. This work, if carefully carried out, should lead to the unambiguous identification of at least some codons, as to composition, order of bases, and size.

The evidence to be obtained from mutation and from genetic studies generally is more difficult to forecast, especially if the code is highly degenerate. Mutations being basically rare events, it follows that arguments based on them are difficult to make completely certain. Does nitrous acid act almost always as we think it does? Do most mutations arise from changes in a single base? Are transversions perhaps very rare? Nevertheless, in time it should be possible, checking the genetic evidence by biochemical studies, to discover how to use it with confidence.

At all stages we may expect evidence on the universality or quasi-universality of the code to accumulate. Any experiment that purports to prove that the code is *not* universal will require careful documentation.

It is obvious that there is no shortage of experiments that need doing. The problem is always which one to try first. Unless the cell-free system shows unexpected difficulties, there is every chance that it can be used to solve the code.

C. On the Place of Theory

It does not seem to be appreciated that theoretical work is often of two rather distinct types. There is first the deduction from experiment: the weighing of the data and the reasoned assessment of, say, the evidence that a particular codon represents a particular amino acid. This I would call interpretation, and it needs to be done critically.

Second we have theory proper. This may take several forms; for example, Wall's demonstration that a *partially* overlapping code is not yet eliminated; or Woese's attempt to deduce the whole structure of the code from only part of it. These theories may not be correct but they are both sensible and useful, in that they enable us to tighten up our logic and make us scrutinize the experimental evidence to some purpose. Moreover, even if Woese's code is wrong, his careful exploration of its consequences may enable us to see something about the general character of the genetic code. But, most important of all, these ideas are not merely useful, they are novel. If their authors had not suggested them, they might not have occurred to many people working on the problem.

The bad theoretical paper takes an obvious technique, applies it to shaky data, and reaches a solution that even on general grounds can be seen to be unlikely. What is gained by this it is difficult to discover.

In the long run we do not want to *guess* the genetic code, we want to

know what it is. It is, after all, one of the fundamental problems of biology. The time is rapidly approaching when the serious problem will be not whether, say, UUC is *likely* to stand for serine, but what evidence can we accept that establishes this beyond reasonable doubt. What, in short, constitutes proof of a codon? Whether theory can help by suggesting the general structure of the code remains to be seen. If the code does have a logical structure there is little doubt that its discovery would greatly help the experimental work. Failing that, the main use of theory may be to suggest novel forms of evidence and to sharpen critical judgement. In the final analysis it is the quality of the experimental work that will be decisive.

Addendum

A fair number of papers have appeared since this review was written, but I shall content myself with mentioning briefly only one or two of the more important ones.

Henning and Yanofsky (1962b) have reported their results on the amino acids substitutions caused by mutation at one particular place in the polypeptide chain of the A protein of the tryptophan synthetase of *E. coli*. In brief, a glycine in the wild type is changed by mutation either to a glutamic acid, from which it can revert to either the original glycine or to valine or alanine; or it is changed to arginine, and can revert to either glycine or serine. In addition they have produced glycine as a *genetic recombinant* of the arginine and glutamic acid mutants, and both serine and glycine as recombinants of the valine and arginine mutants. No recombinants could be detected between the valine and glutamic acid mutants.

These results show clearly that recombination can take place within a codon, though only rarely, as would be expected. The authors note that their results are compatible with Ochoa's triplets—in fact they can be accounted for by a doublet code—but such assignments should be made with great caution until we know the composition of most of the codons. However it is obvious that this type of evidence will eventually be very useful in helping to decide the order of bases within certain codons.

Gardner *et al.* (1962) have now reported the composition of the polynucleotides used previously by Ochoa's group. These deviate rather less than might have been feared from the expected values, the worst case being a poly U,C,G which has base ratios of 6:0.6:1.1 from an incubation mixture with base ratios of 6:1:1. All the polymers containing only two bases had their base ratios within 6% of the expected values.

Wahba *et al.* (1962) have reported preliminary experiments which

suggest that the order of bases in the triplet (UUA) for tyrosine is, in fact, AUU. They also mention that they have evidence that the tyrosine may lie at the C-terminal end of the polypeptide chain, and that cysteine may be GUU. However, the effects produced are small and independent confirmation by more rigorous methods is very desirable.

The major development in the studies on the cell-free system, however, has been the discovery by both Nirenberg's and Ochoa's groups that most polymers without uracil will stimulate amino acid incorporation (Jones and Nirenberg, 1962; Gardner *et al.*, 1962). In particular the latter workers have shown that poly A produces poly-L-lysine, which had been missed previously because it was soluble in the precipitants used. They have also shown that, as would be expected, the product was digested by trypsin but not by chymotrypsin. The new evidence is such as to make all transposible triplet codes unlikely (see Section VII,C,3).

These results are too extensive to be reported here in detail. However, they show clearly that (assuming triplets) the code has considerable degeneracy and that the different codons for one amino acid are related. Thus, one of the main themes of this review has been justified before its publication. It is to be hoped that attention will now be given to the other theme—the necessity for the rigorous experimental proof of each codon.

Acknowledgments

I should like to thank all those who have allowed me to quote their experimental results in advance of publication, in particular Dr. Wittmann, Dr. Tsugita, Dr. Nirenberg, Dr. Ochoa, Dr. Lengyel, and Mr. Bretscher. I am also grateful to my colleagues at Cambridge both for numerous discussions and for removing many infelicities from the manuscript.

References

Arnstein, H. R. V., Cox, R. A., and Hunt, J. A. (1962). *Nature* **194**, 1042.
Basilio, C., Wahba, A. J., Lengyel, P., Speyer, J. F., and Ochoa, S. (1962). *Proc. Natl. Acad. Sci. U.S.* **48**, 613.
Bautz, E. K. F., and Hall, B. D. (1962). *Proc. Natl. Acad. Sci. U.S.* **48**, 400.
Belozersky, A. N., and Spirin, A. S. (1958). *Nature* **182**, 111.
Benzer, S., and Champe, S. P. (1961). *Proc. Natl. Acad. Sci. U.S.* **47**, 1025.
Benzer, S., and Champe, S. P. (1962). *Proc. Natl. Acad. Sci. U.S.* **48**, 1114.
Benzer, S., and Weisblum, B. (1961). *Proc. Natl. Acad. Sci. U.S.* **47**, 1149.
Berg, P., Bergmann, F. H., Ofengand, E. J., and Dieckmann, M. (1961). *J. Biol. Chem.* **236**, 1726.
Berg, P., Lagerkvist, U., and Dieckmann, M. (1962). *J. Mol. Biol.* **5**, 159.
Bishop, J., Leahy, J., and Schweet, R. S. (1960). *Proc. Natl. Acad. Sci. U.S.* **46**, 1030.
Brenner, S. (1957). *Proc. Natl. Acad. Sci. U.S.* **43**, 687.
Bretscher, M. S., and Grunberg-Manago, M. (1962). *Nature* **195**, 283.

Champe, S. R., and Benzer, S. (1962). *Proc. Natl. Acad. Sci. US.* **48**, 532.
Chapeville, F., Lipmann, F., von Ehrenstein, G., Weisblum, B., Ray, W. J., and Benzer, S. (1962). *Proc. Natl. Acad. Sci. US.* **48**, 1086.
Chargaff, E. (1962). *Nature* **194**, 86.
Crick, F. H. C., Griffith, J. S., and Orgel, L. E. (1957). *Proc. Natl. Acad. Sci. US.* **43**, 416.
Crick, F. H. C., Barnett, L., Brenner, S., and Watts-Tobin, R. J. (1961). *Nature* **192**, 1227.
Dintzis, H. M. (1961). *Proc. Natl. Acad. Sci. US.* **47**, 247.
Doctor, B. P., Apgar, J., and Holley, R. W. (1961). *J. Biol. Chem.* **236**, 1117.
Gamow, G. (1954). *Nature* **173**, 318.
Gamow, G., and Yčas, M. (1955). *Proc. Natl. Acad. Sci. US.* **41**, 1011.
Gardner, R. S., Wahba, A. J., Basilio, C., Miller, R. S., Lengyel, P., and Speyer, J. F. (1962). *Proc. Natl. Acad. Sci. US.* **48**, 2087.
Garen, A., and Siddiqi, O. (1962). *Proc. Natl. Acad. Sci. US.* **48**, 1121.
Goldstein, A., and Brown, B. J. (1961). *Biochim. et Biophys. Acta* **53**, 438.
Haschemeyer, A. E. V., and Rich, A. (1962). *Biochim. et Biophys. Acta* **55**, 994.
Helinski, D. R., and Yanofsky, C. (1962). *Proc. Natl. Acad. Sci. US.* **48**, 173.
Henning, U., and Yanofsky, C. (1962a). *Proc. Natl. Acad. Sci. US.* **48**, 183.
Henning, U., and Yanofsky, C. (1962b). *Proc. Natl. Acad. Sci. US.* **48**, 1497.
Hoagland, M. B. (1960). *In* "The Nucleic Acids" (E. Chargaff and J. N. Davidson, eds.), Vol. III, p. 349. Academic Press, New York.
Jacob, F. (1961). *Cold Spring Harbor Symposia Quant. Biol.* **26**, 34.
Jones, O. W., and Nirenberg, M. W. (1962). *Proc. Natl. Acad. US.* **48**, 2115.
Josse, J., Kaiser, A. D., and Kornberg, A. (1961). *J. Biol. Chem.* **236**, 864.
Jukes, T. H. (1962a). *Biochem. Biophys. Research Communs.* **7**, 281.
Jukes, T. H. (1962b). *Biochem. Biophys. Research Communs.* **7**, 497.
Lee, K. Y., Wahli, R., and Barbu, E. (1956). *Ann. inst. Pasteur* **91**, 212.
Lengyel, P., Speyer, J. F., and Ochoa, S. (1961). *Proc. Natl. Acad. Sci. US.* **47**, 1936.
Lengyel, P., Speyer, J. F., Basilio, C., and Ochoa, S. (1962). *Proc. Natl. Acad. Sci. US.* **48**, 282.
Lerman, L. S. (1961). *J. Mol. Biol.* **3**, 18.
Matthaei, J. H., and Nirenberg, M. W. (1961). *Proc. Natl. Acad. Sci. US.* **47**, 1580.
Matthaei, J. H., Jones, O. W., Martin, R. G., and Nirenberg, M. W. (1962). *Proc. Natl. Acad. Sci. US.* **48**, 666.
Maxwell, E. S. (1962). *Proc. Natl. Acad. Sci. US.* **48**, 1639.
Nathans, D., Notani, G., Schwartz, J. H., and Zinder, N. D. (1962). *Proc. Natl. Acad. Sci. US.* **48**, 1424.
Nirenberg, M. W., and Matthaei, J. H. (1961). *Proc. Natl. Acad. Sci. US.* **47**, 1588.
Nirenberg, M. W., Matthaei, J. H., and Jones, O. W. (1962). *Proc. Natl. Acad. Sci. US.* **48**, 104.
Ofengand, E. J., and Haselkorn, R. (1962). *Biochem. Biophys. Research Communs.* **6**, 469.
Perutz, M. F. (1962). "Proteins and Nucleic Acids: Structure and Function." Elsevier, Amsterdam.
Reichmann, M. E., Rees, M. W., and Markham, R. (1962). *Biochem. J.* **84**, 86P.
Rendi, R., and Ochoa, S. (1961). *Science* **133**, 1367.
Roberts, R. B. (1962). *Proc. Natl. Acad. Sci. US.* **48**, 897.
Rolfe, R., and Meselson, M. (1959). *Proc. Natl. Acad. Sci. US.* **45**, 1039.
Rothman, F. (1961). *Cold Spring Harbor Symposia Quant. Biol.* **26**, 23.

Signer, E. R., Torriani, A., and Levinthal, C. (1961). *Cold Spring Harbor Symposia Quant. Biol.* **26**, 31.
Simha, R., and Zimmerman, J. M. (1962). *J. Theoret. Biol.* **2**, 87.
Sinsheimer, R. L. (1959). *J. Mol. Biol.* **1**, 218.
Smith, E. L. (1962a). *Proc. Natl. Acad. Sci. US.* **48**, 677.
Smith, E. L. (1962b). *Proc. Natl. Acad. Sci. US.* **48**, 859.
Speyer, J. F., Lengyel, P., Basilio, C., and Ochoa, S. (1962a). *Proc. Natl. Acad. Sci. US.* **48**, 63.
Speyer, J. F., Lengyel, P., Basilio, C., and Ochoa, S. (1962b). *Proc. Natl. Acad. Sci. US.* **48**, 441.
Sueoka, N. (1961a). *Cold Spring Harbor Symposia Quant. Biol.* **26**, 35.
Sueoka, N. (1961b). *Proc. Natl. Acad. Sci. US.* **47**, 1141.
Sueoka, N., and Yamane, T. (1962). *Proc. Natl. Acad. Sci. US.* **48**, 1454.
Sueoka, N., Marmur, J., and Doty, P. (1959). *Nature* **183**, 1427.
Swartz, M. N., Trautner, T. A., and Kornberg, A. (1962). *J. Biol. Chem.* **237**, 1961.
Tissières, A., and Hopkins, J. W. (1961). *Proc. Natl. Acad. Sci. US.* **47**, 2015.
Tissières, A., Schlessinger, D., and Gros, F. (1960). *Proc. Natl. Acad. Sci. US.* **46**, 1450.
Tsugita, A. (1962). *J. Mol. Biol.* **5**, 284.
Tsugita, A., Fraenkel-Conrat, H., Nirenberg, M. W., and Matthaei, J. H. (1962). *Proc. Natl. Acad. Sci. US.* **48**, 846.
von Ehrenstein, G., and Lipmann, F. (1961). *Proc. Natl. Acad. Sci. US.* **47**, 941.
Wahba, A. J., Basilio, C., Speyer, J. F., Lengyel, P., Miller, R. S., and Ochoa, S. (1962). *Proc. Natl. Acad. Sci. US.* **48**, 1683.
Wall, R. (1962). *Nature* **193**, 1268.
Weinstein, I. B., and Schechter, A. N. (1962). *Proc. Natl. Acad. Sci. US.* **48**, 1686.
Weisblum, B., Benzer, S., and Holley, R. W. (1962). *Proc. Natl. Acad. Sci. US.* **48**, 1449.
Wittmann, H. G. (1962). *Z. Vererb.-Lehre* **93**, 491.
Woese, C. R. (1962). *Nature* **194**, 1114.
Yanofsky, C., Helinski, D. R., and Maling, B. D. (1961). *Cold Spring Harbor Symposia Quant. Biol.* **26**, 11.
Zubay, G. (1962). *Proc. Natl. Acad. Sci. US.* **48**, 894.
Zubay, G., and Quastler, H. (1962). *Proc. Natl. Acad. Sci. US.* **48**, 461.